STOCHASTIC MODELING AND ANALYSIS OF PATHWAY REGULATION AND DYNAMICS

A Dissertation

by

CHEN ZHAO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2012

Major Subject: Electrical Engineering

STOCHASTIC MODELING AND ANALYSIS OF PATHWAY REGULATION AND DYNAMICS

A Dissertation

by

CHEN ZHAO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Edward R. Dougherty
	Ivan Ivanov
Committee Members,	Ulisses Braga-Neto
	Aniruddha Datta
Head of Department,	Costas N. Georghiades

May 2012

Major Subject: Electrical Engineering

ABSTRACT

Stochastic Modeling and Analysis of Pathway Regulation and Dynamics. (May 2012)

Chen Zhao, B.S., Beijing University of Posts and Telecommunications;

M.Eng., Texas A&M University

Co–Chairs of Advisory Committee: Dr. Edward R. Dougherty
Dr. Ivan Ivanov

To understand effectively and treat complex diseases such as cancer, mathematical and statistical modeling is essential if one wants to represent and characterize the interactions among the different regulatory components that govern the underlying decision making process. Like any other complex decision making network, the regulatory power is not evenly distributed among its individual members, but rather concentrated in a few high power "commanders". In biology, such commanders are usually called masters or canalizing genes. Characterizing and detecting such genes are thus highly valuable for the treatment of cancer. We present a Bayesian framework to model pathway interactions, and then study the behavior of master genes and canalizing genes. We also propose a hypothesis testing procedure to detect a "cut" in pathways, which is useful for discerning drugs' therapeutic effect.

Another important task in cancer research is to understand the mechanisms of action (MOA) of cancer drugs. For a new drug, the correct understanding of its MOA is a key step toward its application to cancer treatments. Using the Green Fluorescent Protein technology, researchers have been able to track various reporter genes from the same cell population for an extended period of time. Such dynamic gene expression data forms the basis for drug similarity comparisons. We design an

algorithm that can identify mechanistic similarities in drug responses, which leads to the characterization of their respective MOAs.

Finally, in the course of drug MOA study, we observe that cells in a hypothetical homogeneous population do not respond to drug treatments in a uniform and synchronous way. Instead, each cell makes a large shift in its gene expression level independently and asynchronously from the others. Hence, to study systematically such behavior, we propose a mathematical model that describes the gene expression dynamics for a population of cells after drug treatments. The application of this model to dose response data provides us new insights of the dosing effects. Furthermore, the model is capable of generating useful hypotheses for future experimental design.

To my wife and my parents

ACKNOWLEDGMENTS

A dissertation cannot be finished in a vacuum. I owe my great appreciation to my two co-advisors: Dr. Edward R. Dougherty and Dr. Ivan Ivanov. They gave me great freedom in research and constantly supported and believed in me even during the most difficult times. Dr. Dougherty has often inspired me by his great passion, optimism and integrity for research. Dr. Ivanov has always been patient and supportive for my unorthodox ideas, and he has taught me many things outside school, including my language skills and communication skills.

I am equally grateful for Dr. Michael L. Bittner during my internship in TGen. As a top level biologist, he was willing to spend many hours discussing with me even the most basic questions in biology. His perspective and intuition in drug development have made this work much more pragmatic and meaningful for biology. I would also like to thank Dr. Jianping Hua for the many creative discussions we had during my internship.

I would also like to thank Dr. Aniruddha Datta and Dr. Ulisses Braga-Neto for teaching me several courses and sharing their valuable time on my committee. Also, I would like to thank all the current and past GSP lab members for maintaining a positive learning environment.

Last but not least, I would like to thank my parents for their mental support, and my lovely wife, Siding Liu, for all the little things she has done for me, so that I can concentrate on my research.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

To effectively understand and treat complex diseases such as cancer, mathematical and statistical modeling is essential if one wants to represent and characterize the interactions among the different regulatory components that govern the underlying decision making process. As we know, cell regulation involves control strategies that employ multiple inputs, multiple layers of feedback, and nonlinear decision functions. Owing to the difficulty of modeling and identifying such systems experimentally, historically biologists have concentrated on marginal interaction between signaling molecules to construct signaling pathways. Therefore, effectively utilizing the existing pathway knowledge holds the key for cancer research. In this dissertation, we devote our efforts to pathway based approach to study gene regulation (Chapter II) and gene dynamics (Chapter III and IV).

In the context of gene regulation study, we consider three important subproblems: first, the representation or the modeling of the underlying pathway knowledge; second, characterizing and detecting master genes and canalizing genes in such model; finally, drug intervention effects in such network model. Much of our efforts are concentrated on the second task, since it is believed that master genes or canalizing genes are the "leverage points" in a network and they could serve as potential therapeutic targets.

In the context of gene dynamics study, we mainly focus on analyzing and comparing gene expression dynamic patterns after drug intervention. Such comparisons form the basis for understanding the mechanism of action (MOA) of cancer drugs, and therefore are essential for cancer drug development in general. Furthermore, in

---

The journal model is *IEEE Transactions on Biomedical Engineering*.

the course of analyzing gene expression dynamics, we observe that cells in a hypothetical homogeneous population do not respond to the drug treatment in a uniform and synchronous way. Instead, each cell makes a large shift in its gene expression level independently and asynchronously from the others. Such phenomenon suggests that gene expression should be studied on the single cell level to account for the cell-to-cell variations. A Markov model is proposed to describe gene expression dynamics for a population of cells after drug treatments. We show that such model is useful for understanding dosing effects. Finally, we show that the model is capable of generating useful hypotheses for future experimental design. In the following sections, we explain each topic in detail.

A. Pathway Regulatory Analysis in the Context of Bayesian Networks Using the Coefficient of Determination

One of the most important problems in systems biology is to model gene regulatory networks. Ultimately, the goal is to design proper therapeutic intervention strategies that can slow down, stop or even reverse the progression of tumors. To this end, there have been numerous attempts to model gene regulations, ranging from deterministic to stochastic, using either discrete-time or continuous-time descriptions of gene interactions. Recently, much attention has been devoted to Boolean networks [1] or probabilistic Boolean networks (PBNs) [2], where gene regulation is formed by a set of logic operations and the dynamic behavior of networks can be readily studied in the context of Markov chains. External control policies based on dynamic programming approaches have also been developed to alter the long run behavior of the BN or PBNs, so that network states are more likely to be in the "desirable" or non-cancerous states [3, 4]. However, one inherent disadvantage of using BN or PBNs is that the

number of states grows exponentially with the number of genes in the network, which makes them difficult to study when the number of constituent genes is beyond $20-30$. Furthermore, the inference of PBNs requires a great deal of temporal data [5], which is rarely the case in the contemporary microarray based experiments. To alleviate the complexity problem, researchers have also focused on the state reduction of PBNs [6], and external control strategies on reduced PBNs [7, 8]. Nevertheless, the complexity of PBNs still poses a challenge in practice applications. Moreover, optimal control policy often requires to "flip" a gene according to the corresponding gene activity profile, which is hard to achieve in practice.

Biology is rich of pathway information that is hand curated and refined by generations of biologists. In the most basic model, one can view a pathway as originating with a single regulatory gene (or protein) whose activation initiates a cascade of gene (protein) responses. Since cell regulation involves a decentralized set of interactions among various control agents present within the cell upon receipt of external or internal signals – for instance, activation of a specific gene may require a combination of transcription factors, and translation to the gene product may be affected by post-transcription events – if one views the cascade of activities resulting from the action of a single regulatory gene, both the strength and specificity of subsequent activities in the cascade may be expected to diffuse through subsequent steps in the cascade. As the regulatory effects propagate, they are progressively modified or limited by interactions with other factors modulating transcription. From a modeling perspective, this means that each edge in a pathway has an associated probability and the degree of regulation exerted by the regulatory gene (protein) at the head of the pathway is characterized in terms of these probabilities. In fact, except for the activation probability of the pathway head, each of these probabilities is conditional. Thus, Bayesian networks can serve as a suitable model for a large portion of genetic pathways and

the uncertainty classes associated with them. Hence, in Chapter II, we employ a tree-structured Bayesian network model to represent and characterize the underlying gene regulations. Note that the purpose is not to propose any new gene regulatory network model, but is rather to apply a suitable model which can readily incorporate prior pathway knowledge to study gene regulation. In general, inferring Bayesian network from data is an NP-hard problem [9]. However, in our application, the structure of the model has already been given from the pathway structures. Thus, only the parameters of the model need to be estimated from data – a much simpler task to do.

Once the Bayesian networks are constructed from prior pathway knowledge, the next step is to study and characterize gene regulation in this framework. Like in any other complex decision making networks, the regulatory power is not evenly distributed among its individual members, but is rather concentrated in a few high power "commanders". In biology, such commanders are usually called master or canalizing genes. Biologically, the concept of master genes is not new, however, only until recently, the framework for master genes has been formed mathematically [10], where it utilizes the Coefficient of Determination (CoD), a measure that quantifies the predictability of a "target" variable from a set of "predictor" variables, to detect master genes. In a similar vein, Martins et al. [11] uses a measure called intrinsically multivariate prediction (IMP) power to characterize and detect canalizing gene, where a "target" gene is considered to be canalizing if its "predictors" genes do not predict it well separately, but together, they predict the target gene with high accuracy. While intuitively appealing, both approaches lack a model-based framework to support their findings.

The concept of master genes has been explored by other groups as well [12, 13]. In their approach, a reverse engineering algorithm is used to identify master regulators. In their definition, master genes/regulators are genes whose collective behavior

determines a certain phenotype. To identify such master regulators, an interaction network consisting of hundreds to thousands of genes is estimated from microarray gene expression data using the ARACNe algorithm [14]. Hence, it is possible to associate each transcription factor (TF) with a list A of regulon genes through the inferred interaction network. In the next step, a list B of signature gene markers which distinguish between phenotypes are identified by some statistical method (e.g. t-test or clustering). Next, for each TF, the overlapping score between its regulon gene list A and the signature gene marker list B is calculated, and master regulators are defined by the TFs whose overlapping scores are among the highest. Fig. 1 shows an example of interaction network between the TFs and their targeted mesenchymal signature genes described in [12]. As can be seen, the network is very complicated and contains potentially spurious edges. The method is intuitively attractive, however, there might be several obstacles for such a method to work in practice. First, the task of inferring an interaction network that consists thousands of genes from only ~ 100 of microarray samples is a severely ill-posed inverse problem. Therefore, the inferred interactions can contain a high false positive rate, even for the proposed ARACNe algorithm. Second, the tasking of identifying differentially expressed gene signatures from ~ 100 microarray samples is again a daunting job. In fact, a range of recent study [15–17] has shown that error estimation and feature selection is often intractable in the small sample settings. Thirdly, there is a lack of a coherent, quantitative definition for master genes, plus, ranking based on some overlapping scores can be arbitrary and *ad hoc*. In fact, recently, in a similar feature ranking problem, Zhao et al. [18] has shown that the estimated errors for the top features can be overly optimistic and therefore selection based on top scored features can be misleading. Nevertheless, the proposed method is still able to find master regulators that are subsequently verified by biological assays, indicating the fact that master regulators

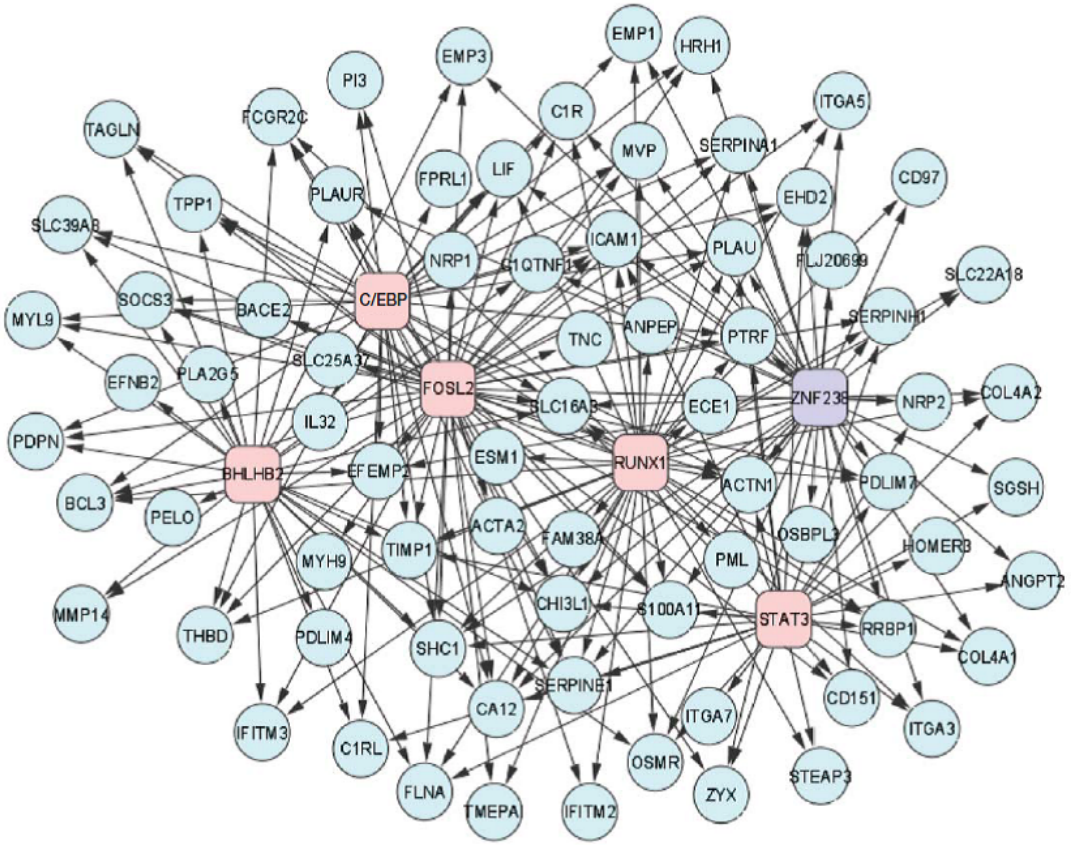


Fig. 1. An interaction network showing the regulation between TFs (pink and purple) and their targeted mesenchymal signature genes (cyan).

exhibit strong control over its slave genes and therefore can serve as potential therapeutic targets. Hence, detecting master genes bears significant biological impacts towards the treatment of cancer.

In Chapter II, we propose a framework to integrate the two concepts introduced by Dougherty and Martins together. Because in both papers, there is a lack of network model to represent the underlying gene-gene interactions, it is unclear what types of nodes will be detected as master genes or canalizing genes and how the different network parameters will affect their detectability. Toward this end, a Bayesian network framework is proposed to represent the underlying pathway infor-

mation with uncertainties. Two measurements called the mean CoD and canalizing power are introduced to detect master genes and canalizing genes respectively. We then conduct a series of simulation studies to examine their relationships with various network structures and parameter values. The results show that both measurements favor "hub" genes; however, the mean CoD approach measures the ability to control while the canalizing power approach measures the ability to take over control. Such subtle difference cannot be appreciated without the help of the proposed network model. Compared to the work in [12, 13], our network model directly comes from prior biological pathway knowledge, and the structure of the network is therefore given, leaving only the parameters of the model to be estimated from data – a much simpler task than structure inference. In the end of Chapter II, we also utilize the mean CoD approach to detect a "cut" in the pathways, which provided a formal statistical approach to discern therapeutic effects.

B. Identifying Mechanistic Similarities in Drug Responses

While it is important to study gene regulations in a probabilistic framework, it is equally important to study gene expression dynamics, since such time series data carries rich information about the evolvement of the underlying signaling network. In the field of cancer drug development, it is often the case that the detailed mechanism of action (MOA) of a drug is not exactly known, since there might be non-specific binding of the drug molecule to the host molecules, various cross-talk and feedback loops to interfere with the drug's effect, etc. Often, the designed drug fails to meet the expected therapeutic effects. To understand the MOA of a drug on a particular cell line, researchers have developed Green Fluorescent Protein (GFP) based method that allows one to track various reporter genes for an extended period of time (up

to 50 hours) [19]. The responses are then summarized by the percentage of cells responded to the drug treatment as well as their averaged fold change. The reporter genes are selected to represent a wide range of the canonical cellular pathways, such as the apoptosis pathways, proliferative pathways, and survival pathways, etc. On the other hand, there are often standard drugs that can attack cell lines at those canonical pathways. Understandably, these various drugs will induce different response characteristics of the cell line, which will then be reflected by the various GFP reporters. Then, we can narrow down the MOA of the new drug by comparing its GFP responses to those of the standard drugs, with the premise that two drugs with similar MOAs should induce similar responses on many of the GFP reporters. As an example, a panel of GFP reporter responses are shown in Fig. 2. Each column corresponds to a drug, while each row corresponds to a GFP reporter, with the gene name shown at the right end of that row. While it is possible to compare reporters responses in a crude way by looking at the general trend of change [19], a quantitative approach is desired to compare drug responses in a more objective and automated way. Hence, there is a need to design proper alignment algorithm to capture the mechanistic similarities among different drugs.

The study of gene expression time series data is not new. The collection of time series gene expression has started from the early days of microarrays. One of the most prominent examples is that of yeast cell-cycle data [20], where the expression levels of 800 genes in yeast rise and fall as the cells go through their reproductive cycle. Khodursky et al [21] measured gene-expression time series in E. coli in order to characterize the genes regulating the bacteria's synthesis of the amino acid tryptophan, in part by observing gene activities over time when the bacteria were exposed to different amounts of external stimuli.

Traditionally, the comparisons of gene expression time series data have been

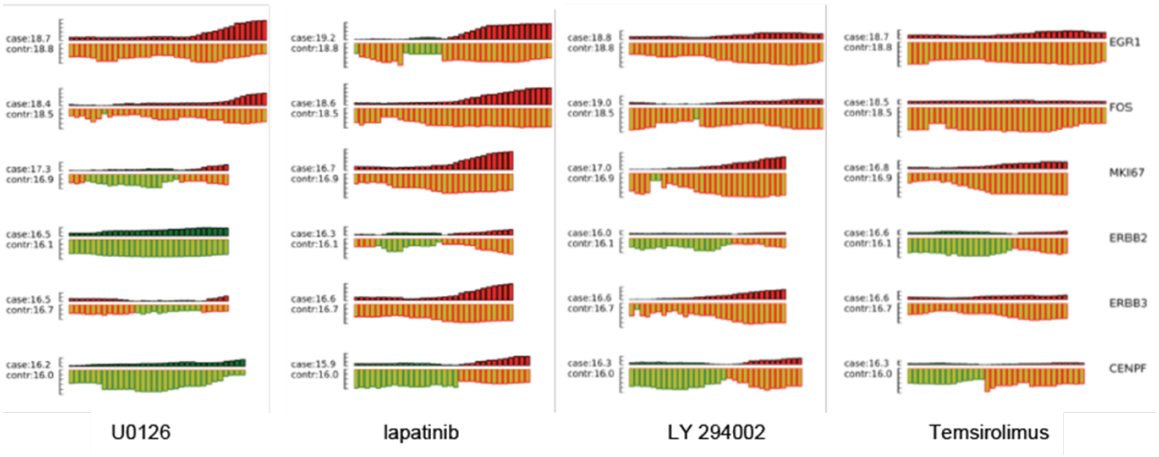


Fig. 2. Gene response dynamics induced by four different drugs on cell line HCT116. The upper panel of each barplot shows the population change and the lower panel shows the corresponding fold change. Red color indicates down-regulation and green color indicates up-regulation.

mostly focused on the so called dynamic time warping (DTW) methods, originally developed by Sakoe and Chiba [22] in the speech recognition community. Aach and Church [23] were the first to apply the method to microarray gene expression profiles, and other groups have followed suit [24, 25]. Briefly, the DTW algorithm works by locally deforming the time axis in order to minimize the cumulative difference between the aligned points. The rationale behind this approach is that biological processes are time elastic, meaning that multiple instances of a single process may unfold at different and possibly non-uniform rates. Therefore, to maximize the similarity, one needs to align them appropriately. Fig. 3 illustrates the type of alignment commonly used in the DTW method.

For several reasons, direct application of DTW type of algorithms is not applicable to the comparison of drug MOAs. First, the goal of DTW is to minimize some cumulative score (usually normalized Euclidean distance), which has no direct biolog-

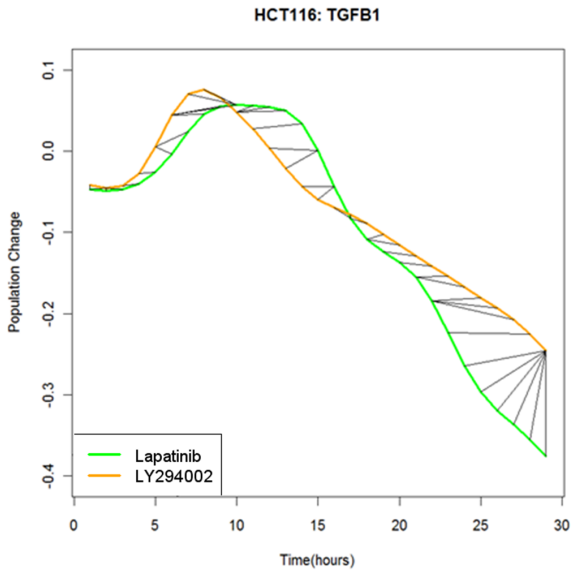


Fig. 3. Schematic figure to show dynamic time warping alignment of two time series data. Time axis is deformed to minimize the cumulative Euclidean distance between the two.

ical meaning. Second, DTW is sensitive to outliers, which can distort the alignment results significantly if present in the signal. Finally, DTW treats all the data points as equally important, however, in the drug response data, the "core" information about the MOA of drugs lies in the region where sufficient cells have responded to the drug. Hence, a good alignment algorithm should bias toward that region to maximize the meaningful biological similarity.

In Chapter III, we propose a recursive method that utilizes the concept of longest common substring idea to iteratively identify biologically interesting similarities in drug response data. The proposed algorithm is able to overcome all of the aforementioned weaknesses of the DTW type algorithm. Applying our algorithm to a range of real drug experiments shows that the newly proposed algorithm is accurate, sensitive and consistent with existing drug MOA knowledge.

C. Modeling Population of Cells' Gene Expression Dynamics after Drug Treatment

Researchers have long realized that gene expression can exhibit a significant degree of variations from cell to cell, even in a hypothetical homogeneous cell population. Many models are proposed to describe such phenomenon. In one possible explanation, the randomness is attributed to the inherent stochasticity in the biochemical process of gene expression (intrinsic noise) or fluctuations in other cellular components (extrinsic noise) [26]. Another theory indicates that gene expression is governed by "transcription bursts", where a gene stays a long time in the inactive state, followed by a short period of active state where it makes a burst of transcripts [27]. Such random bursts lead to different amounts of transcripts inside different cells.

Interestingly, in the course of analyzing gene expression dynamics, we also observe stochasticity among different cells after drug treatments. Briefly, each cell makes a large shift in its gene expression level independently and asynchronously from the others. And the onset response times can vary drastically from cell to cell: some cells respond to the drug very early, but some could respond to the same drug 40 hours later. Intuitively, one can view the drug's effect in any single cell as ineffective or effective. When it is ineffective, a cell stays in its original expression state with a high probability; however, when the drug is effective, it becomes possible for a cell to switch its expression state probabilistically. Therefore, a hybrid system describing the onset response times (modeled as a family of logistic functions) of individual cells as well as the gene expression state transition (modeled as hidden Markov model (HMM)) in each cell is proposed in Chapter IV to describe gene expression dynamics after drug treatment. The model contains key parameters that are biologically relevant, furthermore, we show the model is useful for understanding dosing effect and is capable of generating useful hypotheses for future experimental design.

The use of hidden Markov model has numerous applications, including speech recognition [28], bio-sequence alignment [29, 30], image processing [31, 32], etc. In the field of GFP based cell tracking, Wang et al [33] has recently proposed a HMM based approach to infer cell cycle states from features such as cell shape, size and intensities, etc. However, like in most of the HMM based approaches, their focus is to infer the "correct" hidden states given the observed data. Such problem does not exist in our applications – the hidden gene expression states are directly observable due to the existence of control experiments (see Chapter IV for a detailed description). Therefore, for the model parameter estimation part, we focus on the inference of onset response times, which is critical for understanding dosing effects.

It should be emphasized that our model assumes the gene regulation is governed by two regulation states or regulation contexts – drug ineffective or drug effective. In the two contexts, the associated Markov model has different transition probabilities. A similar contextual regulation idea was previously introduced in [10]. In that paper, the authors assumed that the context is determined by latent variables which lead to probabilistic gene regulations. Compared to their approach, we explicitly model the contexts as drug effects, and more importantly, we consider the gene expression dynamics rather than static gene regulations.

D. Dissertation Outline

The remainder of this dissertation is organized as follows:

- In Chapter II, we first introduce the tree based Bayesian network to model pathway regulations. Two important parameters – cross-talk and conditioning are used to quantify the tightness of regulation between a parent node and its child node. Following the modeling part, we introduce two measurements that

are used to detect master genes and canalizing genes, respectively. We then study the effects of network structure and parameters on the detectability of master genes and canalizing genes. In the end of Chapter II, we formulate a hypothesis testing procedure to detect a "cut" in pathways.

- In Chapter III, we first formulate the gene expression dynamics alignment problem conceptually, and state explicitly what types of mechanistic similarities are important for the understanding of drug MOAs. Then, we introduce a recursive time series alignment algorithm to iteratively identify such similarities in drug response data. Finally, we apply the proposed method to a set of real drug experiments to evaluate its performance.

- In Chapter IV, we introduce a Markov model to describe gene expression dynamics for a population of cells after drug treatment. Then, we discuss how to infer model parameters from both synthetic and real data. The results show that the model is useful for understanding dosing effects.

- In Chapter V, we summarize the main contributions of the work and discuss some future directions of research.

CHAPTER II

PATHWAY REGULATORY ANALYSIS IN THE CONTEXT OF BAYESIAN NETWORKS USING THE COEFFICIENT OF DETERMINATION*

This chapter presents a model based approach to study master genes and canalizing genes. To set the stage, we first introduce the concept of master genes in the context of pathway regulations and give the rationale of using tree structured Bayesian network to model gene regulations. Then, we propose two measurements to quantify master genes and canalizing genes in the network model respectively. Their behaviors are studied systematically by varying the network structures and parameters. In the end, we also propose a hypothesis testing procedure to test a "cut" in the network model, which is potentially useful for discerning drug therapeutic effects.

A. Pathway Knowledge and Bayesian Network

Differentiated cells in a mature organism spend most of their time maintaining a set of activities that either support their own persistence or contribute to the persistence of the organism of which they are a part. In this state, regulation in the cell is mostly fine-tuning and integration of these established activities and does not involve massive shifting of regulatory states. However, when the cell must coordinate the various intermittently used processes required to achieve other particular operations, such as repairing extensive DNA damage that arose as a result of some environmental insult or entering the cell cycle to produce a daughter cell, large changes in regulation are required. As in any system, excursion away from the normal state of processing

*©2011 Journal of Biological Systems Reprinted, with permission, from "Pathway regulatory analysis in the context of Bayesian network using the coefficient of determination" by C. Zhao, I. Ivanov, M. L. Bittner, E. R. Dougherty, 2011, *Journal of Biological Systems*, 19(4):651-682.

may drive the system to a point from which it cannot return to its normal state. In biology one of the largest dangers associated with either loss of the ability to perform a complex corrective action or the loss of the ability to cease operating in a proliferative mode that produces an excess of cells is cancer. To effectively intervene when cells are trapped in pathological modes of operation it is necessary to build models that capture relevant network structure and include characterization of dynamical changes within the system. The model must be of sufficient detail that it facilitates the selection of intervention points where pathological cell behavior arising from improper regulation can be stopped.

Cell regulation involves control strategies that employ multiple inputs, multiple layers of feedback, and nonlinear decision functions. Owing to the difficulty of modeling and identifying such systems experimentally, historically biologists have concentrated on marginal interaction between signaling molecules to construct signaling pathways. In the most basic model, one can view a pathway as originating with a single regulatory gene (or protein) whose activation initiates a cascade of gene (protein) responses. Since cell regulation involves a decentralized set of interactions among various control agents present within the cell upon receipt of external or internal signals – for instance, activation of a specific gene may require a combination of transcription factors, and translation to the gene product may be affected by post-transcription events – if one views the cascade of activities resulting from the action of a single regulatory gene, both the strength and specificity of subsequent activities in the cascade may be expected to diffuse through subsequent steps in the cascade. As the regulatory effects propagate, they are progressively modified or limited by interactions with other factors modulating transcription. From a modeling perspective, this means that each edge in a pathway has an associated probability and the degree of regulation exerted by the regulatory gene (protein) at the head of the pathway is

characterized in terms of these probabilities. In fact, except for the activation probability of the pathway head, each of these probabilities is conditional. Thus, Bayesian networks can serve is a suitable model for a large portion of genetic pathways and the uncertainty classes associated with them. This chapter provides a modeling framework for pathway representation in the context of Bayesian networks and examines several issues.

To illustrate the pathway scenario, consider the Ras pathway model in Fig. 4. Mutant Ras proteins are found in 20-25% of all human tumors and up to 90% in specific tumor types [34]. The Ras protein sits in the middle of a complex signaling cascade and it functions as a binary switch that controls intracellular signaling networks. Once the extracellular signals are received by the receptor proteins located in the cell membrane and passed on to Ras, it then will transmit the signal to three major downstream pathways involved in cell proliferation, differentiation, apoptosis, etc [35]. In a nut shell, the upstream proteins of Ras are summarized by Receptor $\rightarrow$ Shc $\rightarrow$ Grb2 $\rightarrow$ Sos $\rightarrow$ Ras, and the three major downstream pathways of Ras are illustrated by Fig. 4. The Ras pathways form a tree structure, where many branches are downstream from Ras. This structure helps to explain why Ras is powerful as a potent oncoprotein and is able to drive the cell to neoplastic transformations. On the other hand, one would expect that the deregulation of downstream proteins of Ras may possess far less transforming power as compared to Ras. Indeed, the point is illustrated nicely by the mutant B-Raf kinase, the close cousin of Raf, which is also activated by interaction with Ras. These mutations, which are found in many human melanomas, create oncogenic BRAF alleles that have transforming powers that are only about one-fiftieth of those for the activated Ras oncoprotein [36]. Presumably, signaling components located further downstream, when altered by mutation, confer even less transforming power [35]. The model proposed in this chapter provides

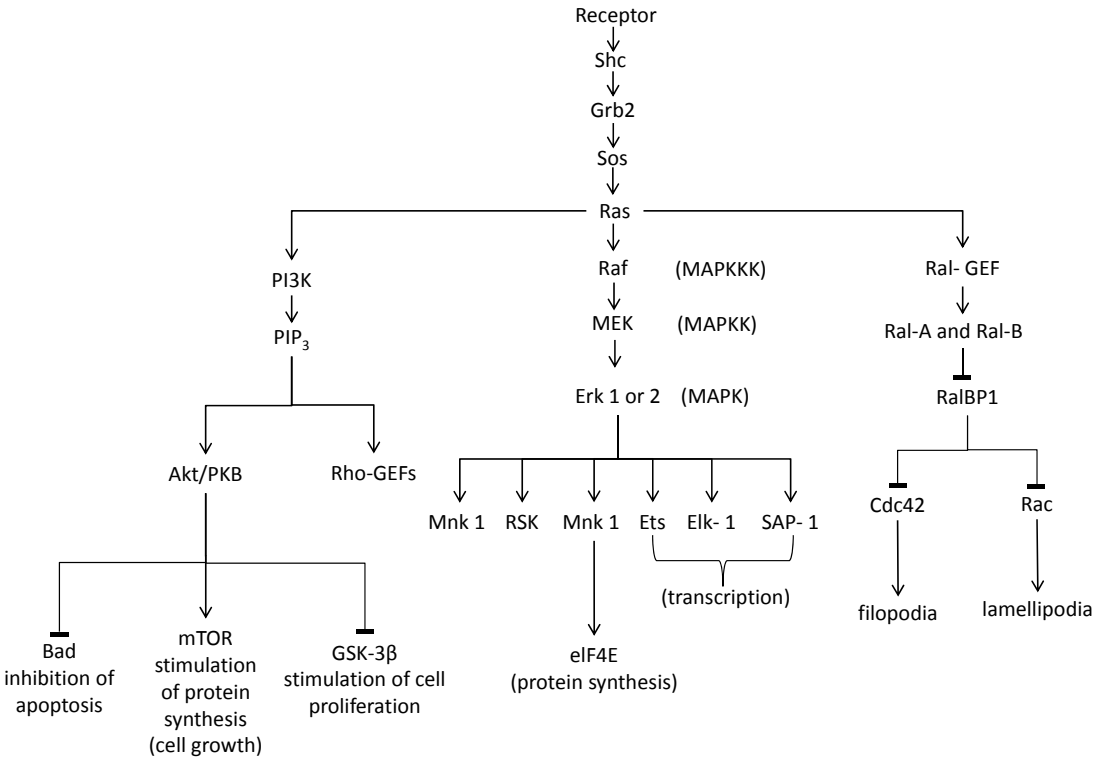


Fig. 4. The Ras protein network.

a mathematical way to characterize these kinds of behaviors, specifically, to model protein-protein (gene-gene) interactions in the framework of graphical models and study their regulatory importance within the model.

This kind of pathway information can play an important role in developing computational methods for identifying and validating drug targets. For instance, in Imoto *et al.* [37], the authors discuss how Bayesian networks can be inferred from microarray data and then used to identify their respective root nodes as the potential regulators or, as the authors term them, "druggable genes". There are several major differences between our work and what is discussed in that chapter: first, our analysis does not

begin with network inference, rather, the pathway structure is given as prior knowledge and the conditional probabilities, i.e., model parameters, are inferred from data by standard statistical techniques; second, we focus on characterizing important nodes (canalizing and master genes) in the model; third, we also discuss statistical testing procedures for detecting pathway disruption. Pathway disruption has previously been considered from a purely logical perspective by treating a set of pathways as a deterministic wiring diagram and then applying classical fault-detection to determine suitable drug combinations [38], but that analysis did not involve any probabilistic considerations. More generally, one could consider various cell-line platforms for the purposes of drug discovery and validation [39]. Such platforms can potentially benefit from our proposed framework that allows for statistical testing of drug disruption of known cell regulatory pathways.

Looking at Fig. 4 from a local perspective, except for the genes at the bottom of the cascade, each gene may be considered as a master for the gene below it, which can be considered as its slave. Moving up a level of perspective, PI3K, Raf, and Ral-GEF may be considered masters for their respective branches, with each gene in a branch being a slave for its respective master. Taking a maximally global perspective, if we consider the sequence Shc $\rightarrow$ Grb2 $\rightarrow$ Sos $\rightarrow$ Ras as a communication channel from the Receptor to Ras, then Ras can be considered as a system master with all other genes in the full downstream pathway being considered its slaves [36]. (From a logical perspective, one could alternatively consider Shc as a system master, in which case Ras would lie within the Shc system.) If gene g is a master for a collection of slave genes, then we would expect that activation of g would be predictable from observation of the slaves and that the wider the swath of control exercised by g, the greater the extent of that predictability as we consider more genes in the network. It is important to emphasize that the concept of a "master" is both relative and local.

A gene that is a "master" for a given portion of a regulatory network could be a "slave" in a different context, e.g., network segment.

The problem considered in Dougherty *et al.* [10] was to quantitatively characterize master genes, more specifically, the power of master genes, via the ability to predict their behavior from the behavior of other genes. Predictive strength was quantified via the *Coefficient of Determination* (CoD), which quantities the increased ability to predict a random variable via a set of "predictor" random variables as opposed to merely predicting it from its own statistics. The model constructed in that paper was purely probabilistic, without any structural considerations. Here we consider the master-slave paradigm in the framework of a Bayesian pathway model. This approach allows a finer characterization of master-slave behavior and allows us to quantitatively characterize regulatory strength in terms of the branching structure of control.

Related to master genes are genes that can constrain, or canalize, a biological system to particular options [40]. We are not referring to sequential canalization, whereby a specific action of the master enforces a cascade of actions among a single highly correlated cohort of genes important in a single process, but rather where a gene has such broad regulatory power, and its action sweeps across such a wide swath of processes, that the full set of affected genes are not highly correlated under normal conditions. Early observations of canalization along the mitogenic pathway involved the Ras gene family, members of which were found to have frequent mutations in their twelfth codon in cancers that produce uncontrolled proliferation [41]. Another significant instance of canalization involves the gene TP53 in regard to stresses to the genome [42]. A key characteristic of a canalizing gene is its ability to override other regulatory instructions if the condition of the cell so warrants. This affects the predictability of the controlling (canalizing) gene by those genes it controls. This property has led to the characterization of canalization via intrinsically multivariate

prediction (to subsequently be defined rigorously), which relates to the ability of a full set of predictors to provide excellent prediction, whereas leaving out any one of the predictors greatly reduces prediction accuracy [11]. In analogy to the master-slave paradigm in Dougherty *et al.* [10], canalization is characterized in Martins *et al.* [11] by the CoD absent a structural model. Here we will characterize canalization in the context of pathways, thereby taking into account structural considerations, and provide a clear discrimination between a master gene and a canalizing gene. Thus, our approach provides a framework, rather than an algorithm, for modeling specific regulatory structures or pathways and their respective uncertainty classes commonly observed in the context of cell regulation. The framework is outlined on Fig. 5, which also emphasizes the relationships among different sections in the chapter.

B. Background

1. Bayesian Networks

Given a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, a Bayesian network $B = (G, \Theta)$ is defined by: (1) a directed acyclic graph (DAG) G whose vertices correspond to X_1, X_2, ..., X_N and (2) a set Θ of local conditional probability distributions for each vertex X_i, given its parents in the graph [43–45]. The graph encodes the "Markov assumption," which states that each variable X_i is independent of its non-descendants, given its parents in G. By the chain rule of probabilities, any joint distribution satisfying the Markov assumption can be decomposed into a product of the local conditional probabilities. Letting $Pa(X_i)$ denote the Markovian parents of X_i in the graph G, the joint probability distribution (JPD) P is completely specified by

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | Pa(X_i)) \tag{2.1}$$

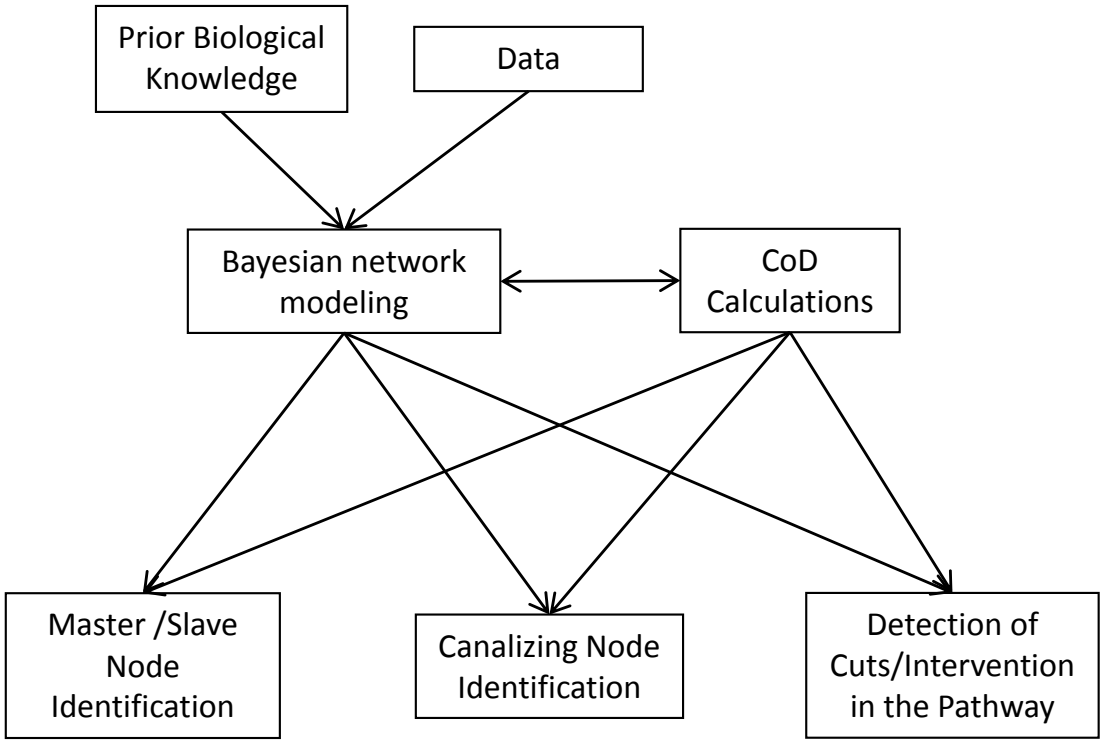


Fig. 5. Framework for modeling pathway regulation showing the relationship between the Bayesian network model, prior knowledge, and inference from data, as well as the application of CoD calculations to detect important nodes or cuts in the pathway.

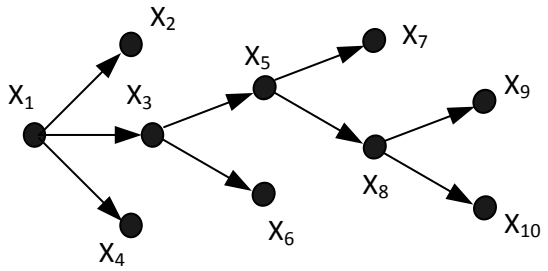


Fig. 6. A tree with the root node X_1.

where $P(X_i|Pa(X_i))$ specifies the conditional probability table (CPT) for X_i and we refer to the parameters that determine $P(X_i|Pa(X_i))$, $i = 1, 2, \ldots, N$, as the parameters of the Bayesian network. Here, we consider the binary case, $X_i = 0$ or 1; however, the results can be easily extended to discrete-valued random variables which assume any pre-specified set of values.

We focus on Bayesian networks, whose DAGs are trees: the DAG has a unique root node with no parents and every other node has precisely one parent and is a descendent of the root [45, 46]. A tree is shown in Fig. 6, the root being X_1.

Once the DAG of a Bayesian network and its associated CPTs are given, the joint distribution is determined by Eq. (2.1) and the joint distribution of any subset of the nodes can be computed from the full joint distribution by summing out the nodes not in the subset. However, this approach is inefficient, since the number of operations grows exponentially with the number of nodes outside the subset. Many efficient inference methods exist for Bayesian networks. It should be noted that Pearl developed a message-passing algorithm for exact inference in Bayesian networks with tree structures. The algorithm proceeds by passing two types of messages among neighboring nodes iteratively and computing the conditional probabilities of interest by updating the two types of messages [47]. Using the message passing algorithm, we

Table I. CPT of X_i in the tree model.

$P(X_i = 0 \mid X_j = 0) = 1 - \eta_{ji}$	$P(X_i = 1 \mid X_j = 0) = \eta_{ji}$
$P(X_i = 0 \mid X_j = 1) = \delta_{ji}$	$P(X_i = 1 \mid X_j = 1) = 1 - \delta_{ji}$

can compute the joint distribution of any subset of variables in the tree efficiently.

The parameters of a tree are specified as follows. Let X_r be the root of the tree. Let $C_{r,0} = P(X_r = 0)$ and $C_{r,1} = P(X_r = 1)$. If X_j is the parent of X_i, then the CPT of X_i is given by Table I, where δ_{ji} is the "conditioning parameter" between X_j and X_i and its magnitude depends on the extent to which the influence of X_j on X_i is diminished by contextual effects and η_{ji} is the "cross-talk parameter" and its magnitude depends on the effects of other nodes during the periods when the parent X_j is not actively regulating X_i [10]. Intuitively, if both δ_{ji} and η_{ji} are small, then the regulation between X_j and X_i should be tight; conversely, if both δ_{ji} and η_{ji} are large, then the regulation should be loose. The marginal distribution of any node is easily computed once the parameters of the tree are given. We let $C_{i,0} = P(X_i = 0)$ and $C_{i,1} = P(X_i = 1)$.

Our interest is to model signaling pathways with uncertainties in gene regulation. Although every node, except the root, in the tree has only one parent, this does not mean that each gene has only one physical regulator in the actual pathway. In fact, X_i could have multiple regulators, X_j, X_{j+1} , ... , X_{j+q}, and this regulation could change in different cell contexts. By focusing on the relation between X_j and X_i, the regulation appears random rather than deterministic and we use the cross-talk and conditioning parameters to capture the uncertainties. In this sense, the tree model is a higher level abstraction of gene regulation rather than a model of the detailed physical interactions among different genes. It is important to recognize that

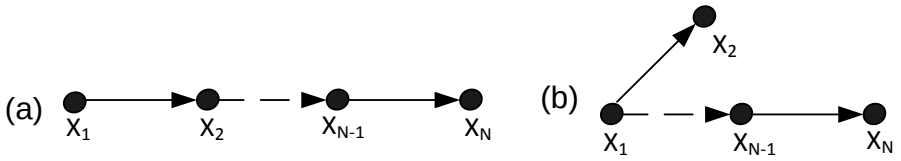


Fig. 7. Two basic types of trees.

the overall effect of cross-talk and conditioning depend on where they occur in the pathway. Furthermore, the structure of the model (that is, nodes and edges) is not restricted only to gene-gene interactions. It can also be used to model more general relationships between various factors participating in cell regulatory pathways, for example, signaling molecules and/or proteins represented by network nodes and their probabilistic relationships represented by network edges.

A salient attribute of Bayesian networks is their ability to encode conditional independencies among variables, which reduces significantly the number of parameters required to represent a complex joint distribution and facilitates efficient probabilistic computations. In fact, given the DAG of a Bayesian network, we can directly read properties of conditional independencies between random variables. For example, for the tree in Fig. 7(a), given X_2, X_1 is independent of X_i, for $i = 3, \ldots, n$; indeed, $P(X_1|X_2) = P(X_1|X_2, X_3, \ldots, X_n)$. Intuitively, the information flow from X_1 to X_i is blocked once X_2 is known. For Fig. 7(b), given X_1, X_2 is independent X_i, for $i = 3, \ldots, n$; indeed, $P(X_2|X_1) = P(X_2|X_1, X_3, \ldots, X_n)$ [46].

2. Coefficient of Determination

The *Coefficient of Determination*(CoD) measures the relative decrease in error when optimally predicting a random variable X using random vector $\mathbf{Y}$ as opposed to optimally predicting X based only on its own statistics. Formally, the CoD for $\mathbf{Y}$

predicting X is defined in Dougherty *et al.* [48] by

$$CoD_{\mathbf{Y}}(X) = \frac{\varepsilon_0(X) - \varepsilon_\bullet(X, \mathbf{Y})}{\varepsilon_0(X)} \tag{2.2}$$

where $\mathbf{Y}$ is a random vector composed of r "predictor" random variables, which are used to predict the status of the "target" node X, $\varepsilon_0(X)$ is the mean-square error (MSE) for predicting X using only its own distribution, and $\varepsilon_\bullet(X, \mathbf{Y})$ is the minimal MSE for using $\mathbf{Y}$ to predict X, which means that it is the minimal error achieved over all possible functions that predict the value of X from the values of the components of $\mathbf{Y}$ [48]. When using the CoD we are not interested in any particular function of $\mathbf{Y}$ that predicts X; rather, we are only concerned with the performance of the optimal prediction of X based on $\mathbf{Y}$. It is in this way that the CoD determines the inherent strength of the connection between a target gene and its predictors. The CoD measures nonlinear interaction and is therefore more appropriate to genomics than the correlation coefficient, which only measures linear interaction.

The CoD has been used since the early days of microarray analysis to characterize the nonlinear multivariate interaction between genes, where the problem was to utilize expression measurements to determine whether or not the expression level of one gene can be predicted by the values of others, with gene expression quantized to three levels: 1 (up-regulated), -1 (down-regulated), and 0 (invariant) [49]. The CoD measures nonlinear association (increase in prediction power), not causality. When $CoD_{\mathbf{Y}}(X)$ is high, it does not indicate that the set $\mathbf{Y}$ of genes regulates X (directly or indirectly); instead, it could mean that X regulates the random variables composing $\mathbf{Y}$ (directly or indirectly). Indeed, herein, the CoD will be used to measure the strength of downstream genes predicting upstream genes. The intuition is that, if gene g regulates genes g_1 and g_2, the observation of g_1 and g_2 should allow one to predict the behavior of g, the stronger the control by g, the stronger the prediction.

Owing to its ability to quantify the degree of interaction, the CoD has been used for numerous purposes in genomics, including: iteratively growing gene regulatory networks from seed genes by adjoining new genes strongly connected to those currently included in the growing network [50], characterization of canalizing genes [11], the identification of master genes [10, 51], and the reduction of gene regulatory networks for the purpose of lowering computational complexity while at the same time preserving regulatory information [7]. Since, in practice, the CoD is typically estimated from data, error estimation performance is a key issue and performance comparison among commonplace estimation procedures (resubstitution, cross-validation, and bootstrap) has been studied [52].

We restrict ourselves to the binary case (0 and 1); however, the basic definition for $CoD_{\mathbf{Y}}(X)$ is not so restricted. In the binary setting, there are simple expressions for $\varepsilon_{\bullet}(X, \mathbf{Y})$ and $\varepsilon_0(X)$; Letting y_1, y_2 ,..., y_{2^r}, denote the 2^r possible values for $\mathbf{Y}$, running from $(0, 0, \ldots, 0)$ to $(1, 1, \ldots, 1)$,

$$\begin{aligned} \varepsilon_0(X) &= min\{(X = 0), p(X = 1)\} \\ \varepsilon_{\bullet}(X, \mathbf{Y}) &= \sum_{j=1}^{2^r} min\{P(X = 0, \mathbf{Y} = \mathbf{y}_j), P(X = 1, \mathbf{Y} = \mathbf{y}_j)\} \end{aligned} \tag{2.3}$$

where the computation of $\varepsilon_{\bullet}(X, \mathbf{Y})$ requires the joint distribution of X and $\mathbf{Y}$.

Note that conditional independencies are important in calculating CoDs. For example, in Fig. 7(a), $CoD_{X_2,X_3}(X_1) = CoD_{X_2}(X_1)$ because $P(X_1|X_2, X_3) = P(X_1|X_2)$. Intuitively, X_3 is blocked by X_2 and does not provide any additional information in predicting X_1.

C. CoD and Basic Tree Structures

To understand the relationships between the CoD, the cross-talk and conditioning parameters, and the tree structure, we start by investigating two basic structures in a tree. The three-node chain in Fig. 8(a) possesses the JPD in Table II. The marginal probability distribution for X_1 and X_3 is given in Table III.

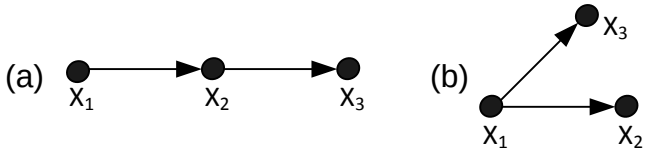


Fig. 8. Two three-gene trees.

Table II. Joint probability distributions of a 3-gene chain shown in Fig. 8(a).

X_1	X_2	X_3	JPD
0	0	0	$C_{1,0}(1-\eta_{12})(1-\eta_{23})$
0	0	1	$C_{1,0}(1-\eta_{12})\eta_{23}$
0	1	0	$C_{1,0}\eta_{12}\delta_{23}$
0	1	1	$C_{1,0}\eta_{12}(1-\delta_{23})$
1	0	0	$C_{1,1}\delta_{12}(1-\eta_{23})$
1	0	1	$C_{1,1}\delta_{12}\eta_{23}$
1	1	0	$C_{1,1}(1-\delta_{12})\delta_{23}$
1	1	1	$C_{1,1}(1-\delta_{12})(1-\delta_{23})$

Consider the special case when $\eta_{12} = \delta_{12} = \eta_{23} = \delta_{23} = a$. Then $\eta_{13} - \eta_{12} = \delta_{13} - \delta_{12} = a - 2a^2 > 0$ for $0 < a < 0.5$. The regulation becomes weaker as the signal

Table III. Marginal probability distributions for the 3-gene chain shown in Fig. 8(a) : $\eta_{13} = (1-\eta_{12})\eta_{23} + \eta_{12}(1-\delta_{23})$ and $1-\delta_{13} = \delta_{12}\eta_{23} + (1-\delta_{12})(1-\delta_{23})$.

X_1	X_3	JPD
0	0	$C_{1,0}(1-\eta_{13})$
0	1	$C_{1,0}\eta_{13}$
1	0	$C_{1,1}\delta_{13}$
1	1	$C_{1,1}(1-\delta_{13})$

propagates along the pathway, the diminishing regulation depending on a, as shown in Fig. 9. Note that $CoD_{X_3}(X_1) < CoD_{X_2}(X_1)$, since X_1 loses its control along the path with increased cross-talk and conditioning. As in Fig. 8(a), $CoD_{X_2,X_3}(X_1) = CoD_{X_2}(X_1)$. This relationship does not result from the specific choice of the cross-talk and conditioning parameters; it is solely determined by the independencies encoded in the 3-gene chain.

The JPD for the 3-gene branch in Fig. 8(b) is given in Table IV. If we assume the same parameter settings as in the 3-gene chain, then $CoD_{X_2}(X_1) = CoD_{X_3}(X_1) = CoD_{X_2,X_3}(X_1)$, as shown by the dashed line in Fig. 9. Considering a more interesting example in Fig. 10, we let $\delta_{12} = \delta_{13} = b$, with $0 \leq b \leq 0.5$, we fix $\eta_{12} = \eta_{13} = a = 0.5$, and let $C_{1,0} = 0.5$. The intention of this choice is to see the behaviors of CoD values of X_1 when the cross-talk is high and the conditioning varies from low to high. We see that $CoD_{X_2,X_3}(X_1) > CoD_{X_2}(X_1)$ in Fig. 10. Intuitively, because X_2 and X_3 are in different branches originating from X_1, they provide complementary information about X_1, thereby resulting in an increase in prediction power.

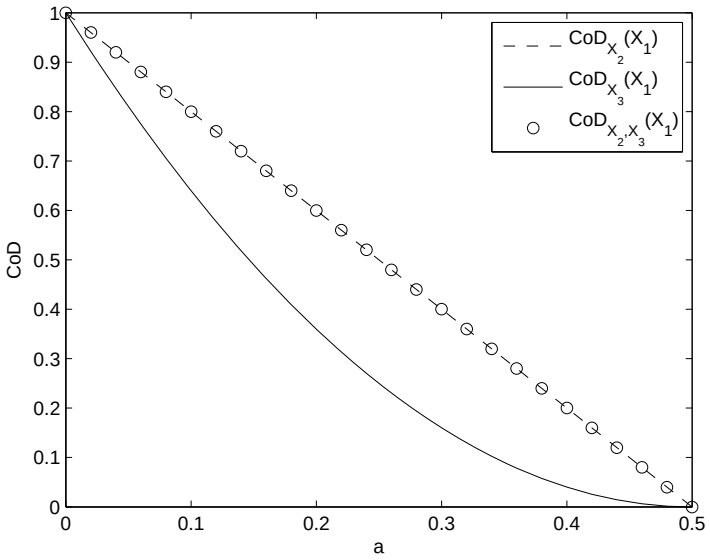


Fig. 9. CoDs for a 3-gene chain shown in Fig. 8(a): $C_{1,0} = 0.5$ and $\eta_{12} = \delta_{12} = \eta_{23} = \delta_{23} = a$. Note that $CoD_{X_3}(X_1) < CoD_{X_2}(X_1)$, because X_1 loses its control power over X_3 along the path with increased cross-talk and conditionings.

D. Master/Slave Paradigm

1. Master and Slave Genes in the Context of a Bayesian Network

The master-slave paradigm explores the relationship between the CoD histograms of a particular gene and its regulatory importance [10]. The CoD histogram for a particular gene is generated by computing CoD values by all possible predictor sets of size r. For example, if there are a total of 10 genes of interest and if we consider all possible pairs of predictors, the CoD histogram of any given gene should include $C(9, 2) = 36$ CoD values. It was hypothesized that CoD histograms skewed to the right (high mean CoDs) correspond to master genes and CoD histogram skewed to the left (low mean CoDs) correspond to slave genes [10]. This interpretation is based

Table IV. Joint probability distributions of a 3-gene branch shown in Fig. 8(b).

X_1	X_2	X_3	JPD
0	0	0	$C_{1,0}(1-\eta_{12})(1-\eta_{13})$
0	0	1	$C_{1,0}(1-\eta_{12})\eta_{13}$
0	1	0	$C_{1,0}\eta_{12}(1-\eta_{13})$
0	1	1	$C_{1,0}\eta_{12}\eta_{13}$
1	0	0	$C_{1,1}\delta_{12}\delta_{13}$
1	0	1	$C_{1,1}\delta_{12}(1-\delta_{13})$
1	1	0	$C_{1,1}(1-\delta_{12})\delta_{13}$
1	1	1	$C_{1,1}(1-\delta_{12})(1-\delta_{13})$

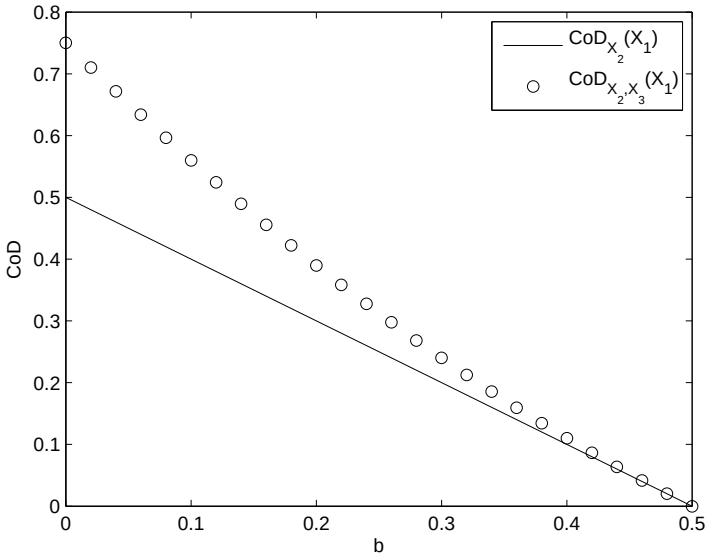


Fig. 10. CoDs for a branch shown in Fig. 8(b): $\delta_{12} = \delta_{13} = b$, with with $0 \leq b \leq 0.5$, $\eta_{12} = \eta_{13} = a = 0.5$, and $C_{1,0} = 0.5$. Note that $CoD_{X_2,X_3}(X_1) > CoD_{X_2}(X_1)$, because X_2 and X_3 can provide complementary information about X_1.

on the observation that if a master gene potentially regulates many other slave genes, then many of the pairs formed by those slave genes should serve as good predictors for the master, thereby producing high CoDs. In this chapter, rather than relying on a general interpretation of the CoD, we examine the matter in the framework of a tree model representing regulatory pathways. In particular, we relate the mean CoD values of a gene to its regulatory importance in the model.

It should be noted that the concept of master/slave model was originally proposed using the Boolean formalism, where the output of the Boolean functions may vary depending on particular contexts or hidden variables [10]. The notions of cross-talk and conditioning were introduced to incorporate these uncertainties. In the current Bayesian-network framework the source of uncertainty may include this interpretation but is not limited to it.

The mean CoD of a node X_i using all single predictors in a tree dis given by

$$CoD_S(X_i) = \frac{\sum_{j=1, j\neq i}^{N} CoD_{X_j}(X_i)}{N-1} \tag{2.4}$$

The mean CoD of a node X_i using all double predictors in a tree is given by

$$CoD_D(X_i) = \frac{\sum_{1\leqslant j<k\leqslant N, j,k\neq i} CoD_{X_j,X_k}(X_i)}{C(N-1,2)} \tag{2.5}$$

Eq. (2.5) gives the average strength of predicting X_i by using all of the possible pairs of the rest of the genes in the network. Intuitively, if X_i is a master gene, then, its activity should be able to influence many of its slave sets and therefore, $CoD_D(X_i)$ should be high [10].

Consider the tree in Fig. 11, where for illustration purposes we assume common cross-talk and conditioning parameters for each node. We plot the mean CoD of each node as a function of the two parameters and visualize the changes directly.

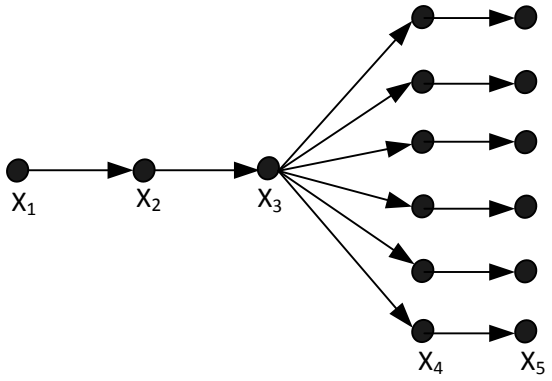


Fig. 11. A tree with 5 layers and 15 nodes. Due to symmetry, only one representative gene on each layer is annotated, assuming common cross-talk and conditioning parameters for each node.

Furthermore, we can compare the mean CoD plots for different genes and see how they reflect their respective regulatory importance.

$CoD_D(X_i)$ for each X_i and two-gene prediction is plotted in Figs. 12 and 13 as a function of η and δ, assuming $C_{1,0} = 0.5$ and $C_{1,0} = 0.1$ in Figs. 12 and 13, respectively – The brighter the image, the higher the mean CoD values. We observe: (1) The node X_3 has the highest mean CoDs, indicating that the "hub" gene is likely to be detected as the master gene using CoDs; (2) the node X_5 has relatively low mean CoDs, indicating that the downstream gene is likely to be detected as a slave gene by the CoD approach; and (3) when $C_{1,0} = 0.1$, $CoD_D(X_1)$ is extremely low, which occurs because $\varepsilon_0(X_1) = 0.1$ and therefore it is hard to achieve an increase of prediction by other genes. Were there no cross-talk or conditioning between the root and the hub, then they would be equivalent relative to $CoD_D(X_i)$ and therefore the root would also be detected.

Having observed that the mean CoD approach tends to detect hub genes in the tree model, we now investigate how the mean CoD changes with respect to the number

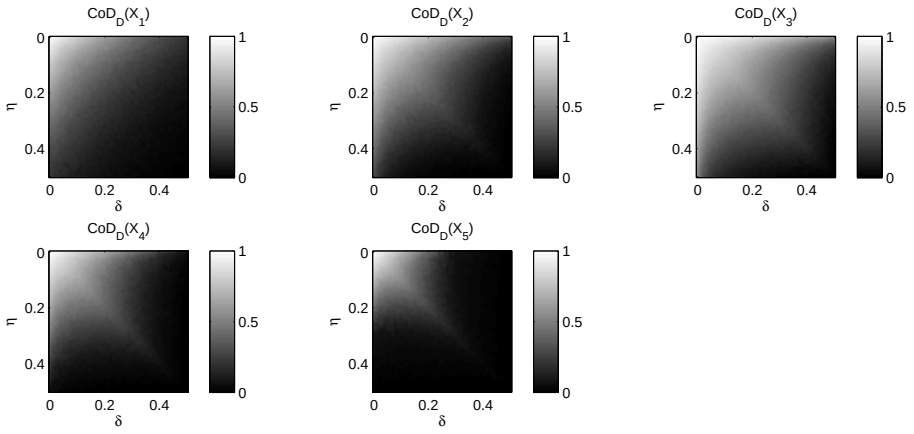


Fig. 12. Plots of $CoD_D(X_i)$ corresponding to the tree in Fig. 11, with $C_{1,0} = 0.5$.

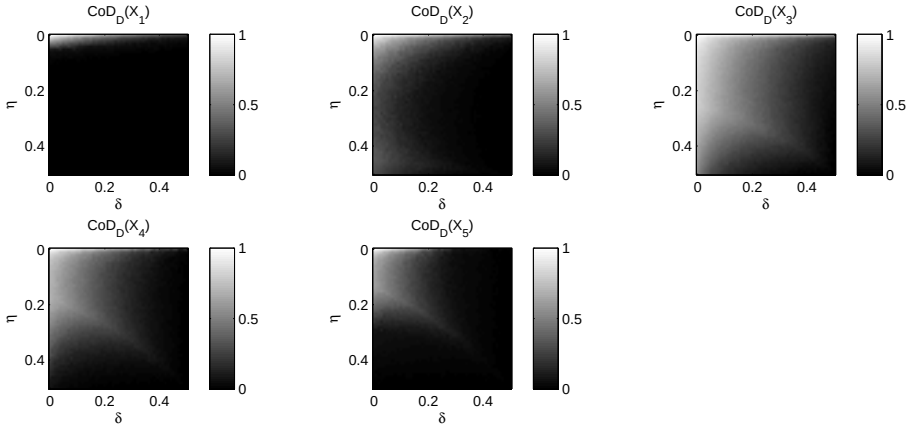


Fig. 13. Plots of $CoD_D(X_i)$ corresponding to the tree in Fig. 11, with $C_{1,0} = 0.1$.

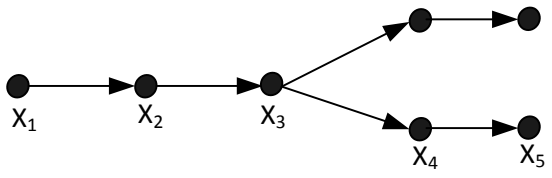


Fig. 14. A tree with 5 layers and 7 nodes.

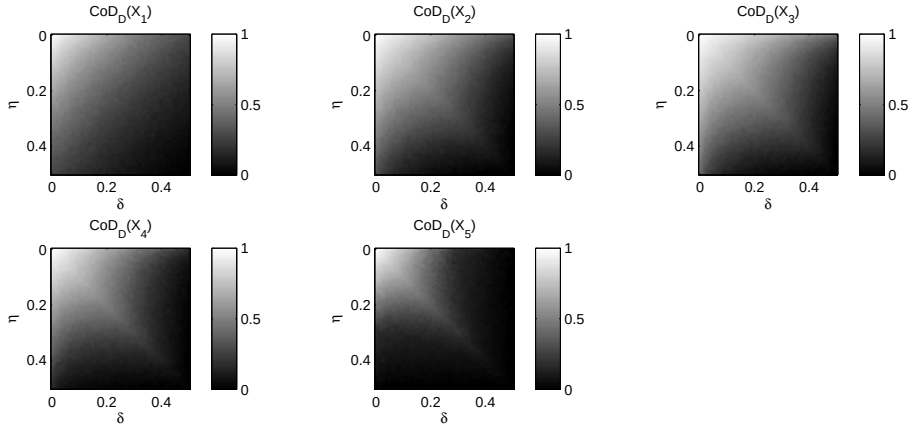


Fig. 15. Plots of $CoD_D(X_i)$ corresponding to the tree in Fig. 14, with $C_{1,0} = 0.5$.

Table V. Comparisons for overall double mean CoD intensities for X_1 through X_5 in two different trees. The values for the 15-node tree and 7-node tree represent the overall intensities of each gray scale image in Figs. 12 and 15, respectively.

	$CoD_D(X_1)$	$CoD_D(X_2)$	$CoD_D(X_3)$	$CoD_D(X_4)$	$CoD_D(X_5)$
15-node tree	641.49	736.76	1006.1	597.73	350.52
7-node tree	826.78	903.78	994.1	771.86	496.94

of branches in the model. Consider the tree in Fig. 14, which is similar to the tree in Fig. 11 except that the hub gene X_3 has only 2 branches going out. $CoD_D(X_i)$ for each X_i is plotted in Fig. 15 as a function of η and δ, assuming $C_{1,0} = 0.5$. Since less outgoing branches suggests less genes controlled by that gene, we expect $CoD_D(X_3)$ to be lower in Fig. 14 in comparison to Fig. 11. Table V confirms this expectation. In particular, the hub gene X_3 stands out more from the rest of the genes in the 15-node tree because it controls more branches.

2. An Example of a TP53 Pathway

In this section we construct a TP53 pathway to illustrate the pathway methodology being developed. The NCI 60 ACDS is a set of widely studied human cancer cell lines derived from cancers of colon, breast, ovary, lung, kidney, prostate, central nervous system, skin, and bone marrow. In the original study, duplicate cultures of 64 cell lines were either irradiated with a high dose of ionizing radiation and harvested four hours later or left untreated and harvested four hours later. From the entire data set, 40 cell lines containing an equal number of TP53 positive and TP53 negative members were chosen to allow investigation of radiation response in cells. The data are binarized and, from an original set of 496 genes, a subset, A, of 96 genes is kept eliminating those with variance not exceeding 0.19 (see [10] for more details on the data set). We obtain a cell cycle, cancer and cell death network generated by Ingenuitiy (http://www.ingenuity.com/). This network is curated purely based on expert knowledge and contains a total of 35 genes (not shown). 16 of these 35 belong to A. If we focus on the connections between TP53 and the other 15 genes, we obtain a simplified tree structure whose root is TP53 (Fig. 16). Although the proposed simplification could lead to loss of information embedded in the original network, it aims to capture the regulatory power of TP53.

To illustrate how the tree model and the proposed mean CoD approach can be used for the purposes of master gene characterization, we performed the following experiment.

1. Given the tree structure Bayesian network on Fig. 16, we estimated the associated network parameters from the 40 samples. Specifically, for each node in the tree, the conditional probabilities of it being ON or OFF given its parent was estimated using the Bayesian estimation approach with $Beta(1, 1)$ prior [46].

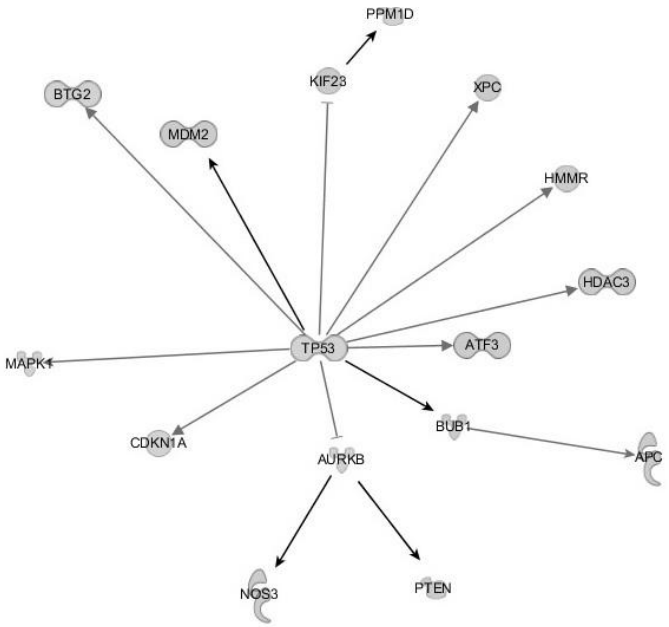


Fig. 16. A tree model for TP53.

2. For each node in the tree, we drew values for the cross-talk and conditioning parameters from their respective posterior Beta distributions, see step 1. Once the parameters were drawn for all of the nodes in the tree, we calculated the single-gene mean CoD for each node in the tree, as shown in Eq. (2.4).

3. We repeated the previous step 1000 times, so that for each node, 1000 single-gene mean CoDs were calculated. The histograms for the 6 representative genes from Fig. 16 are shown on Fig. 17.

4. We also calculated the empirical single-gene CoD for each gene directly from the 40 samples (red crosses shown on Fig. 17). The empirical CoD calculations were done similar to what is described in Section C of this chapter, except that the joint probability distributions were replaced by the maximum likelihood estimations calculated from the 40 samples. Interested readers should refer

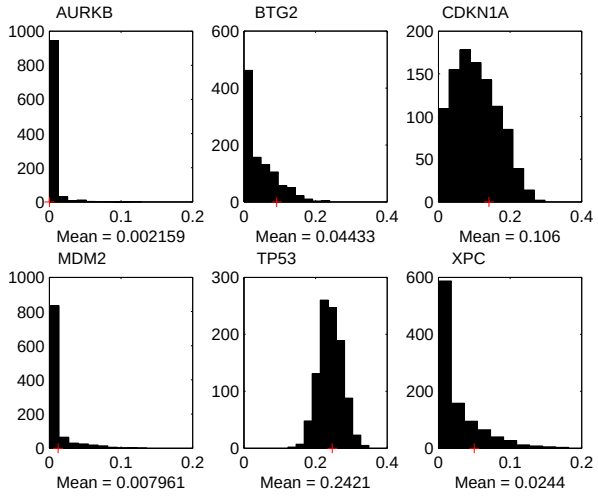


Fig. 17. Single-gene mean CoD distributions for 6 representative genes in Fig. 16. The red-cross indicates the empirical-mean CoD computed directly from the 40 samples. Note that the mean CoD of TP53 stands out from the rest of the genes, indicating its role as a master regulator.

to [52] for a detailed study of various CoD estimators.

This example shows that the model can capture the regulatory power of TP53 and the behavior of the other 15 genes in the sense that all of the empirical-mean CoDs are within the 0.95 confidence interval of the mean CoDs obtained from the tree model, i.e., the red crosses in Fig. 17 agree well with the means of their respective histograms. Moreover, TP53 shows up as the master gene, in the sense that it has the biggest single-gene mean CoD.

E. Canalizing Genes

The first quantitative study of canalization was in the context of logical functions in the Boolean network framework [1, 53–55]. A canalizing function is one in which at least one of the input variables has one value that is able to determine the value of the

output of the Boolean function, regardless of the other variables [55]. For example, the Boolean function $Z = X \vee Y$ is a canalizing function, since $X = 1$ implies $Z = 1$, regardless of the value of the input variable Y. There is also evidence that many control rules governing transcription of eukaryotic genes are canalizing when viewed in the Boolean formalism [56]. The preceding definition of a canalizing function attempts to characterize canalizing genes locally from the perspective of Boolean logic; however, it does not characterize the role of canalizing genes from a network perspective, i.e. globally. In this section, we define and study the canalizing properties of a gene in the proposed tree model. As explained later, such definition favors genes that are directly connected to multiple branches in the model, and therefore have the potential to take over the control of many pathways.

1. Canalizing Gene Definition in the Tree Model

Previously, Intrinsically Multivariate Predictive (IMP) scores were used to detect canalizing genes [11]. The IMP score for gene X_i is defined as

$$\Delta^i_{j,k} = CoD_{X_j,X_k}(X_i) - \max(CoD_{X_j}(X_i), CoD_{X_k}(X_i)) \tag{2.6}$$

where $\Delta^i_{j,k}$ is the increase in prediction power using two predictors over the maximum of the two CoDs when using each predictor individually. The IMP score quantifies the synergistic prediction effect of the pair. The definition is naturally extended to more genes but we will not need that here. We define the canalizing power, $t_N(X_i)$ of a gene X_i, relative to the tree model (G, Θ), by

$$t_N(X_i) = \sum_{1 \le j < k \le N, j,k \ne i} \Delta^i_{j,k} \tag{2.7}$$

The canalizing power measures the total increase in prediction power using pairs of predictors over the maximum of the respective single predictors. As N grows, the

Table VI. CPT of X_2 and X_3 given X_1 in Fig. 8(b), with $C_{1,0} = 0.5$, $\eta_{12} = \eta_{13} = \eta = 0.5$, and $\delta_{12} = \delta_{13} = \delta = 0$.

$P(X_2 = 0, X_3 = 0 \mid X_1 = 0)$	0.25
$P(X_2 = 0, X_3 = 1 \mid X_1 = 0)$	0.25
$P(X_2 = 1, X_3 = 0 \mid X_1 = 0)$	0.25
$P(X_2 = 1, X_3 = 1 \mid X_1 = 0)$	0.25
$P(X_2 = 1, X_3 = 1 \mid X_1 = 1)$	1

canalizing power $t_N(X_i)$ will also grow. The canalizing power of a gene is quantified by the extent of the synergistic prediction from all genes in the model.

To illustrate the above definition, we consider the 3-gene branch shown in Fig. 8(b). To facilitate intuition of Eq. (2.7), we assume that $C_{1,0} = 0.5$, $\eta_{12} = \eta_{13} = \eta$, $\delta_{12} = \delta_{13} = \delta$, and $0 \leq \delta \leq \eta \leq 0.5$. Straightforward calculations yield $t_3(X_1) = \delta^2 - \delta + \eta - \eta^2$, $t_3(X_2) = 0$ and $t_3(X_3) = 0$. For fixed cross-talk η, $t_3(X_1)$ is a parabola that is a decreasing function on $\delta \in [0, 0.5]$. The maximum value is attained when $\delta = 0$ and $\max(t_3(X_1)) = 0.25$ when $\eta = 0.5$. $t_3(X_2) = t_3(X_3) = 0$, since $P(X_3|X_1, X_2) = P(X_3|X_1)$ and $P(X_2|X_1, X_3) = P(X_2|X_3)$, X_1, X_2 and X_3 being conditional independent from each other. Table VI gives the CPT of X_2 and X_3 given X_1 when $C_{1,0} = 0.5$, $\eta_{12} = \eta_{13} = \eta = 0.5$, and $\delta_{12} = \delta_{13} = \delta = 0$. It says that when X_1 is OFF, X_2 and X_3 will be equally likely to be in the states $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$; however, when X_1 is ON, X_2 and X_3 can only be in the state $(1, 1)$, i.e., X_1 has taken over the control of X_2 and X_3. This precisely describes the properties of canalizing genes.

2. Canalizing Power and Network Size

To see the effect of increasing the number of nodes in the network shown in the 3-gene branch, let us consider the two possible ways of adding a new node to the branch shown in Fig. 18. For parts (a) and (b), we have

$$t_N(X_1) = \sum_{1<j<k\leqslant N} \Delta^1_{j,k} = t_{N-1}(X_1) + \sum_{1<j<N} \Delta^1_{j,N},$$

and

$$t_N(X_1) = \sum_{1<j<k\leqslant N} \Delta^1_{j,k} = t_{N-1}(X_1) + \Delta^1_{3,N}.$$

respectively. In part (a), the newly added node X_N is able to a form synergistic pair with any node X_i, $i = 2, 3, \ldots, N-1$, whereas in part (b), the newly added node X_N can only form a synergistic pair with X_3. Therefore, we expect to see that the canalizing power of X_1 is much higher in part (a) than the canalizing power of X_1 in part (b). Fig. 19 shows the canalizing power for X_1 in Fig. 18(a) as a function of network size. The power grows faster when there are large discrepancies between the cross-talk and conditioning parameter. Fig. 20 shows the canalizing power for X_1 in Fig. 18(b) as a function of network size. The canalizing power grows slowly and eventually becomes saturated as N becomes larger.

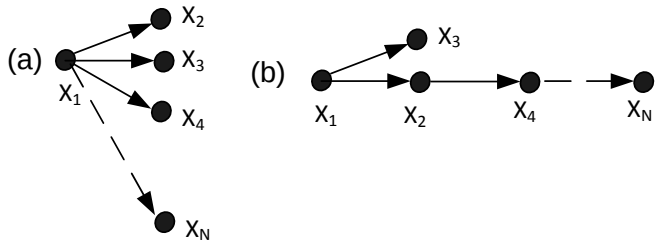


Fig. 18. Adjoining a branch: (a) grow a node directly from X_1; (b) grow a child directly from X_2.

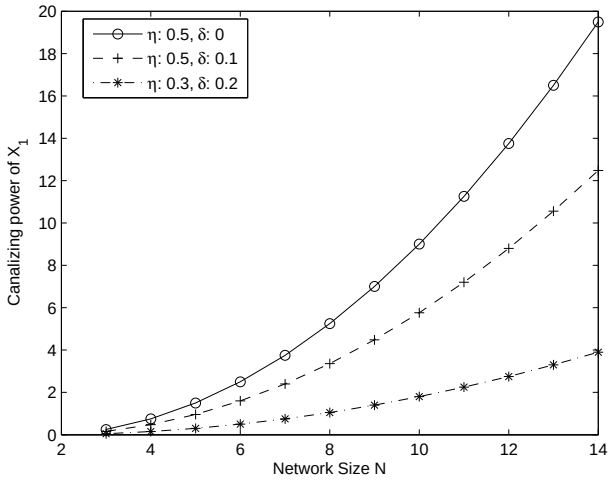


Fig. 19. Canalizing power for X_1 in Fig. 18(a), with $C_{1,0} = 0.5$.

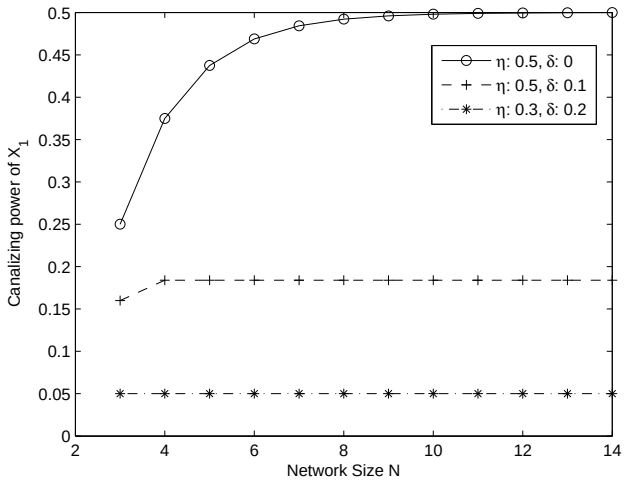


Fig. 20. Canalizing power for X_1 in Fig. 18(b), with $C_{1,0} = 0.5$.

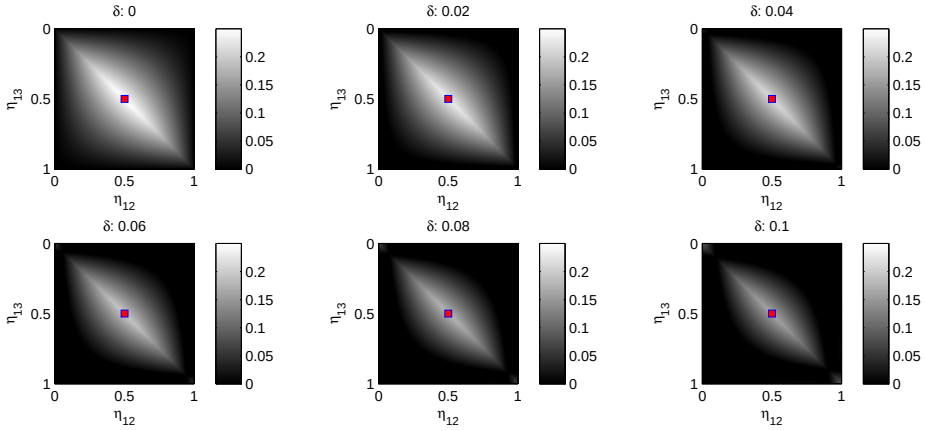


Fig. 21. Canalizing power for X_1 in the 3-gene branch shown in Fig. 8(b), with $C_{1,0} = 0.5$.

3. Canalizing Power and Network Parameters

The joint distribution of the 3-gene branch in Fig. 8(b) is determined by 5 parameters: $C_{1,0}$, η_{12}, η_{13}, δ_{12} and δ_{13}. We can plot the canalizing power of X_1 with respect to η_{12}, η_{13} (or any other two parameters) given the remaining 3 parameters. Fig. 21 shows the canalizing power for X_1 in the 3-gene branch shown in Fig. 8(b), with $C_{1,0} = 0.5$. The intensity represents the canalizing power. The canalizing power decreases as the conditioning parameter $\delta_{12} = \delta_{13} = \delta$ increases. The red square indicates the point with the maximum canalizing power. Fig. 22 shows the canalizing power for X_1 in the 3-gene branch shown in Fig. 8(b), with $\delta_{12} = \delta_{13} = 0$. The maximum canalizing power (red square) increases as $C_{1,0}$ increases.

Fig. 23 is an enlargement of the last subfigure in Fig. 22. Let us focus on the red square in Fig. 23, which corresponds to $C_{1,0} = 0.9$, $\delta_{12} = \delta_{13} = 0$ and $\eta_{12} = \eta_{13} = 0.111$. Table VII shows that, given the status of X_2 and X_3, X_1 is highly predictable. In fact, Table VII resembles the Boolean function $X_1 = X_2 \wedge X_3$. Therefore, X_2 and X_3 has high synergistic prediction power for X_1.

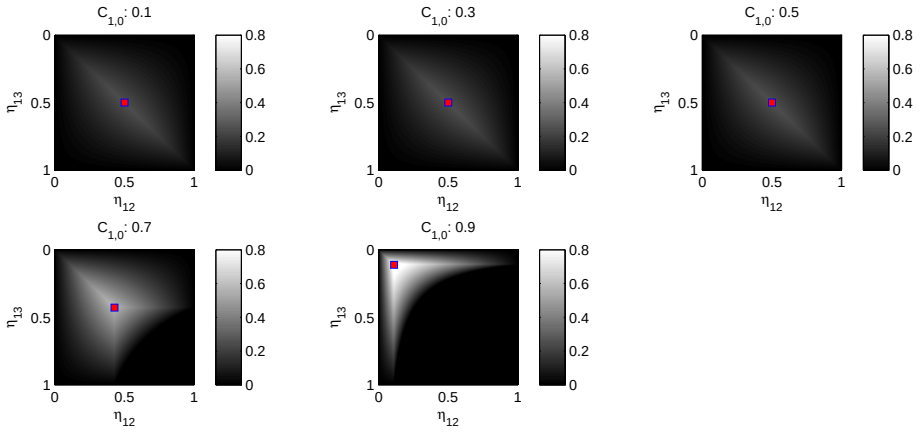


Fig. 22. Canalizing power for X_1 in the 3-gene branch shown in Fig. 8(b), with $\delta_{12} = \delta_{13} = 0$.

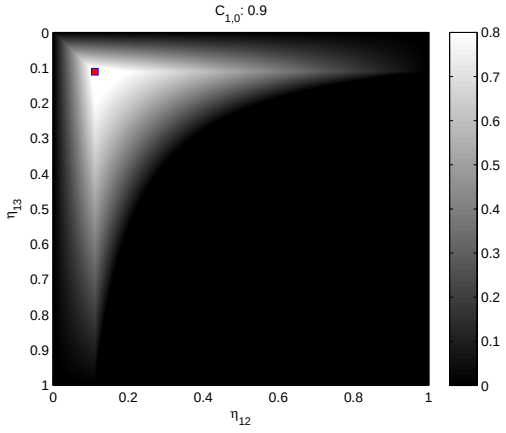


Fig. 23. The last subfigure in Fig. 22. Red square: $C_{1,0} = 0.9$, $\delta_{12} = \delta_{13} = 0$ and $\eta_{12} = \eta_{13} = 0.111$.

Table VII. CPT of X_1 given X_2 and X_3 for the red square in Fig. 23.

$P(X_1 = 0 \mid X_2 = 0, X_3 = 0)$	1
$P(X_1 = 0 \mid X_2 = 0, X_3 = 1)$	1
$P(X_1 = 0 \mid X_2 = 1, X_3 = 0)$	1
$P(X_1 = 1 \mid X_2 = 1, X_3 = 1)$	0.9

It is important to distinguish between master genes and canalizing genes. In our simulations, we see that both definitions tend to reward genes which control many pathways; however, the two concepts are not the same. In fact, for a master gene X_i, we evaluate it by all the possible pair-predictors formed by the rest of the genes in the network. Hence, $CoD_D(X_i)$ is maximized when there is no cross-talk and conditioning in the network and $CoD_D(X_i) = 1$ in this case. On the other hand, for a canalizing gene X_j, we evaluate it by all the possible synergistic pair-predictors formed by the rest of the genes in the network. Thus, if there is no cross-talk and conditioning in the network, $t_N(X_j) = 0$. The reason is that single predictors have already given perfect predictions for X_j and there is no room for improvement for double predictors. Figs. 21 and 22 confirm this observation by showing that in all regions where cross-talk and conditioning are 0, the canalizing powers are also zero. In sum, the master gene definition measures the ability of control, whereas the canalizing gene definition measures the ability of taking over control.

4. An Example of a DUSP1 Pathway

The example given in this section not only shows how the calculations of the canalizing power defined in Eq. (2.7) can be performed on data obtained from a microarray experiment but it also serves as a validation of the ability of the concept of canalizing

power to quantify the known important biological role of some genes, i.e., DUSP1.

We focus on a pathway involving DUSP1 and Ras genes that are especially important in melanoma tumors. The regulatory pathway presented on Fig. 24 is constructed from canonical pathway knowledge. The dataset used to infer the model parameters was obtained in a microarray experiment involving 31 melanoma patient samples, 19 normal tissues and 12 from tissues diagnosed with melanoma. Gene expressions were binarized to indicate change or no change relative to a reference expression level for each gene individually. A change can be under- or over-expression. Both cases are labeled as 1, whereas no significant change from the reference is labeled as 0 [11].

When DUSP1 is OFF, or down-regulated, the downstream (relative to the depicted pathway) genes are controlled by the Ras oncogene through phosphorylation (+p) and transcriptional activation (+Tr). When DUPS1 is ON, or up-regulated, it de-phosphorylates ERK1/2, thereby overriding the signal sent by Ras. The biological role of DUSP1 indicates that it is likely a canalizing gene that can take over control of downstream genes when it is ON. Thus, we expect to capture its behavior by the IMP score and the value of canalizing power as defined in the previous subsection. To test whether DUSP1 shows high canalizing power, we have considered each gray gene from Fig. 24 as a target and computed its respective canalizing power using all possible triple predictors from the rest of the genes measured by the microarrays along the pathway,

$$t_N(X_i) = \sum_{1 \leqslant j<k<l \leqslant N, j,k,l \neq i} \Delta^i_{j,k,l},$$

where

$$\Delta^i_{j,k,l} = CoD_{X_j,X_k,X_l}(X_i) - max(CoD_{X_j}(X_i), CoD_{X_k}(X_i), CoD_{X_l}(X_i)).$$

The results are summarized in Table VIII. Note that DUSP1 significantly stands out from the rest of the genes and the results agree with our knowledge about the biological role of DUSP1.

F. Hypothesis Testing to Detect a "Cut" in the Pathway

Detecting structural changes in a regulatory pathway is critically important when designing therapeutic strategies, e.g. how a drug affects gene regulation in a pathway? CoD measures gene-gene interactions and therefore provides a means to detect a structural change. In the setting of the tree model, a cut between the parent X_j and the child X_i weakens the regulation of X_j on X_i. Fig. 25(a) shows an original tree with a cut tree in Fig. 25(b), the cut occurring between the first and second nodes. Given a cut between parent X_j and child X_i, X_i will be more susceptible to influence from other genes and both the cross-talk η_{ji} and the conditioning δ_{ji} parameters will increase. Zero cross-talk and conditioning indicate deterministic control and 0.5 cross-talk and conditioning indicate no control at all, i.e., given the status of X_j, X_i is equally likely to be ON or OFF. A cut can be partial, meaning it increases the cross-talk and conditioning parameters, but not necessarily to 0.5, or complete, in which case they are increased to 0.5.

When a drug is applied with the intent of cutting the pathway between X_1 and X_2 as shown in Fig. 25(b), we would like to determine its effectiveness. This determination corresponds to two competing hypotheses:

H_0: There is no cut between the first and second nodes in the pathway (drug ineffective).

H_1: There is a cut between the first and second nodes in the pathway (drug effective).

The corresponding quantitative hypothesis test is given by

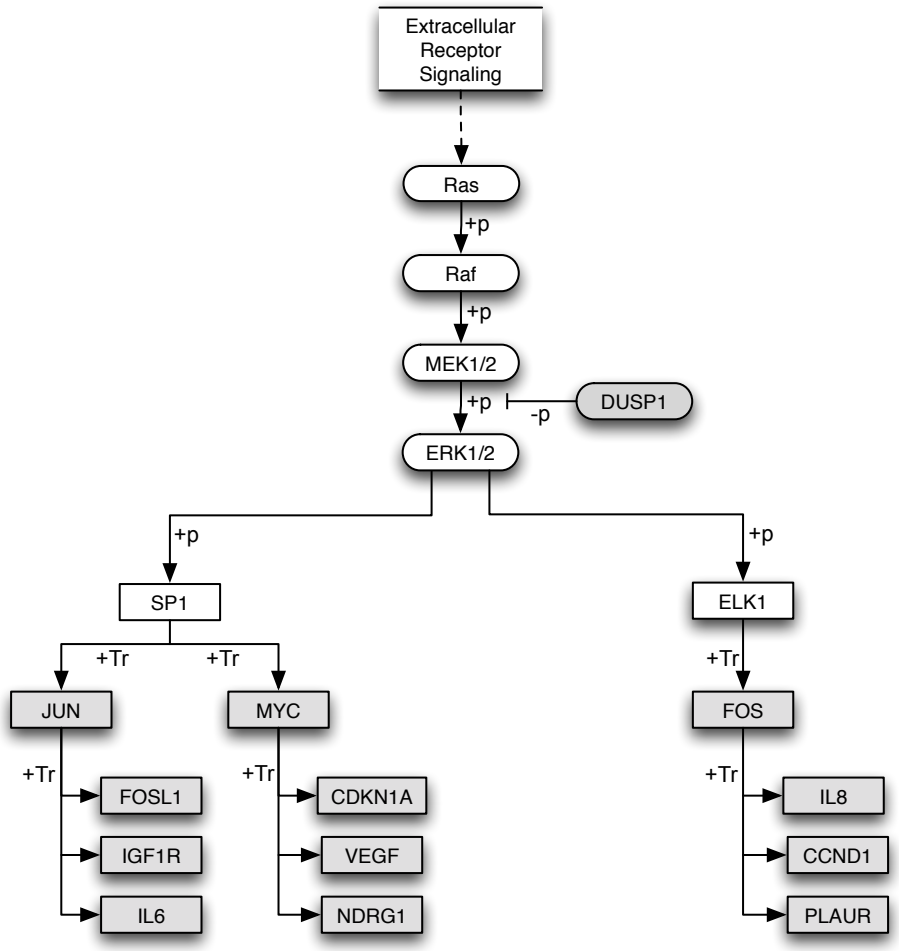


Fig. 24. DUSP1 network: +p indicates phosphorylation, -p indicates dephosphorylation, and +Tr indicates transcriptional activation. The gene expression levels of the gray-colored nodes were measured in an experiment performed at the Translational Genomics Research Institute (unpublished work). There was no measurement of the gene expression levels for the white-colored nodes. Data from experiments show that when turned ON, DUSP1 exerts strong control over the downstream genes via de-phosphorylation of ERK1/2.

Table VIII. Canalizing power for each gray-colored gene in Fig. 24. Canalizing power of DUSP1 stands out from the rest. Only $\Delta^i_{j,k,l} \geq 0.4$ are considered when computing canalizing power for each gene. Note that the order of the list need not follow their topological order depicted on Fig. 24.

Gene Names	$t_N(X_i)$
DUSP1	13.4286
NDRG1	5.1
VEGF	5.0833
JUN	4.6667
CDKN1A	3.6667
IL6	2.6429
MYC	2.5385
IL8	1.7778
IGF1R	1.5
FOSL1	1.2
CCND1	0.9231
FOS	0.8667
PLAUR	0

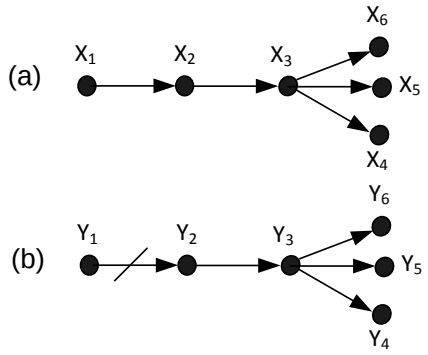


Fig. 25. Original tree (a) and the tree with a cut between the first and second node (b).

H_0: $CoD_S(X_1) \geq T$

H_1: $CoD_S(X_1) < T$

where the test statistic is the empirical-mean CoDs using all single predictors computed from sample data. $CoD_S(X_1)$ represents how strongly X_1 is connected to other nodes in the pathway.

To evaluate the hypothesis test we need the distribution of the test statistic (empirical-mean CoD). Since we do not have an analytic form for the distribution, as we would, let's say, in the case of testing the mean of a Gaussian distribution, we take Monte Carlo approach to generating the distribution. This requires sampling from the pathway, for which we need to know the joint probability distribution of the underlying Bayesian network. Once the JPD is known, it is straightforward to sample from it. For example, in the case of Fig. 25(a), with $C_{1,0} = 0.5$, $\eta_{ji} = 0.1$, and $\delta_{ji} = 0.1$, $1 \leq i < j \leq 6$, we can first generate K samples of X_1, with $P(X_1 = 1) = 0.5$, and then generate X_2 samples based on the probabilities $P(X_2 = 1|X_1 = 0) = 0.1$ and $P(X_2 = 1|X_1 = 1) = 0.9$. In this fashion, we can generate K samples for all 6 nodes in the tree and calculate the empirical mean CoDs using all single predictors from

the K samples. Every time we generate K samples from the pathway, we will get a different empirical mean CoD, thereby forming an empirical-mean CoD distribution.

$T = CoD_S(X_1)$ is calculated when the null hypothesis is true. In the standard way, this is done under the conservative assumption that $CoD_S(X_1) = T$. For example, when $C_{1,0} = 0.5$, $\eta_{ji} = 0.1$, and $\delta_{ji} = 0.1$, $1 \leq i < j \leq 6$ in Fig. 25(a), $CoD_S(X_1) = 0.5949$. T is a population parameter that represents our belief of the *status quo* (drug ineffective).

Fig. 26 illustrates the empirical-mean CoD distribution using single predictors for X_1 under the null hypothesis (solid line), $\eta_{12} = 0.1$ and $\delta_{12} = 0.1$, and under a specific alternative hypothesis (dashed line), $\eta_{12} = 0.2$ and $\delta_{12} = 0.2$, where the empirical CoDs are calculated from 100 samples. The dashed line is shifted more towards the left compared to the solid line, after the cut, the prediction of X_1 from other nodes in the tree is weakened. Under the null hypothesis, we calculate the critical point for the 0.05 significance level to be 0.4882. Under the alternative hypothesis, we calculate the corresponding type II error to be 0.2644. Since the type II error depends on the specific parameter values assumed under the alternative, we can plot type II errors as a function of the alternative values and produce the corresponding operating characteristic curves [57]. These are shown Fig. 27 for sample sizes $K = 50$, 100, and 200 respectively. Note that the type II errors decrease quickly with increasing η_{12} and δ_{12} in all 3 cases. Also, the type II error is smaller for larger sample size K, in other words, it is easier to detect the cut with larger sample size.

To apply the hypothesis test in practice, we first need to estimate the parameters (cross-talk and conditioning parameters) of the pathway. We can take gene expression data from N cell lines and estimate these parameters. Once the tree/pathway is specified (therefore, its JPD), we can generate mean CoD histograms using the same techniques described above. The goal is to generate mean CoD histograms when no

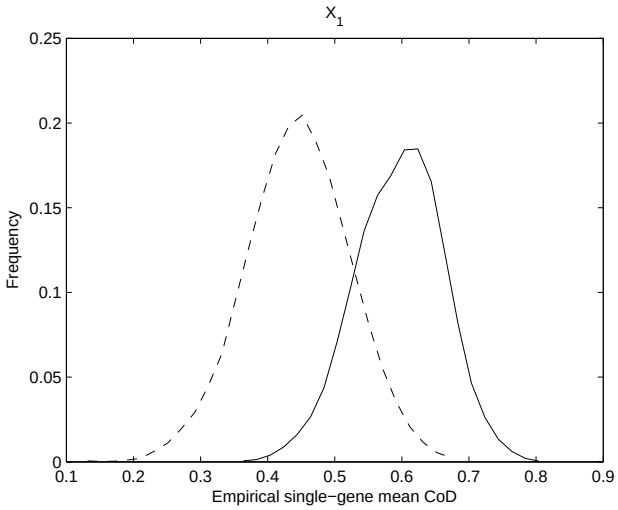


Fig. 26. Empirical mean CoD (single predictor) distribution for X_1 in Fig. 25, with sample size $K = 100$. Solid line for: pathway in Fig. 25(a), with $\eta_{12} = 0.1$ and $\delta_{12} = 0.1$; dashed line for pathway in Fig. 25(b), with $\eta_{12} = 0.2$ and $\delta_{12} = 0.2$. The critical point for 0.05 significance level is 0.4882 and Type II error is 0.2644.

drug is applied. Now, we can apply the drug to M identical cell lines and measure their gene expressions. We can then compute the empirical mean CoD (test statistic) from the M samples. If this test statistic is very small and unlikely to happen under the null distribution generated in the previous step, then, we can claim that the drug is effective and quantify it by a *p-value*.

G. Conclusion

In this chapter we have modeled gene biological pathways in the context of Bayesian networks whose DAGs are trees and examined the relations between CoDs and the tree model extensively. Three regulatory issues have been addressed in this framework: master genes, canalizing genes, and cutting pathways. Our interest in this problem

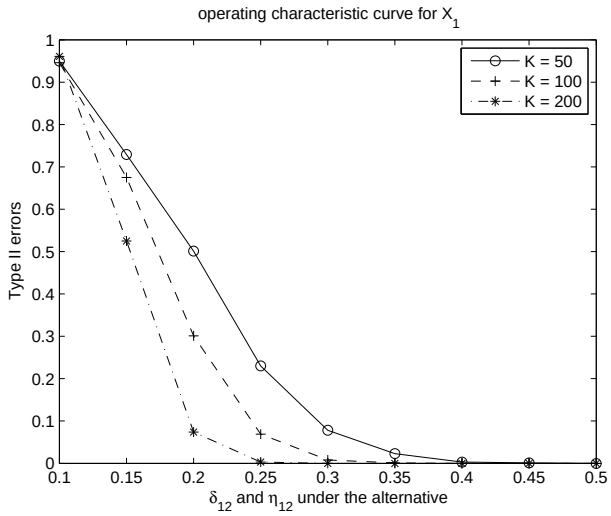


Fig. 27. Operating characteristic curves for X_1 in Fig. 25, with sample size $K = 50$, 100, and 200. Assuming $C_{1,0} = 0.5$, $\eta_{ji} = 0.1$ and $\delta_{ji} = 0.1$, under the null hypothesis.

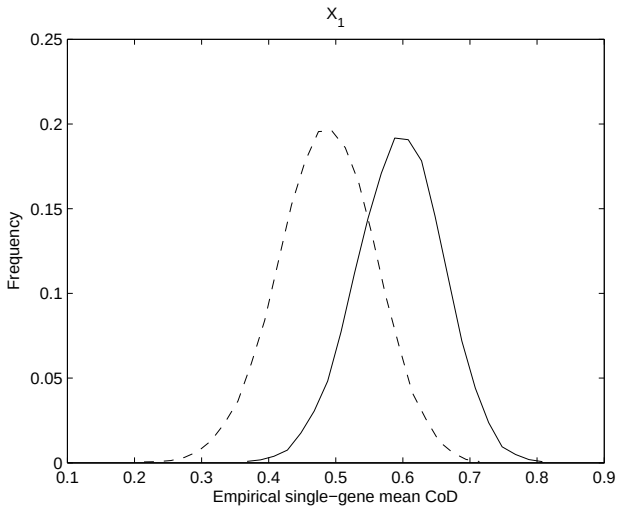


Fig. 28. Empirical mean CoD (single predictor) distribution for X_1 in Fig. 25, with a cut between X_2 and x_3 instead of X_1 and X_2. Sample size $K = 100$: solid line for pathway before the cut, with $\eta_{23} = 0.1$ and $\delta_{23} = 0.1$; dashed line for pathway after the cut, with $\eta_{23} = 0.2$ and $\delta_{23} = 0.2$. The critical point for the 0.05 significance level is 0.4882 and Type II error is 0.4858.

stems from the manner in which regulation dysfunction leads to cancerous phenotypes and our desire to better characterize and mitigate that dysfunction.

Although we have focused on the tree model, the ideas presented can be extended to polytrees, where multiple parents may exist for the same node, albeit, with greater complexity. For instance, Pearl's algorithm can also be extended to compute the joint distribution of any nodes in a polytree [58] and Rebane and Pearl have provided an efficient algorithm to recover polytrees from data [59]. The difficulty is that a node in the polytree with k parents requires 2^k parameters to define the conditional probability table (assuming binary data). Given the limited sample size in a typical genomic experiment, it may be better to stay with simple trees rather than polytrees; however, when a sufficient sample size is available or prior knowledge strongly indicates multiple parents, we may switch to the polytree model.

CHAPTER III

IDENTIFYING MECHANISTIC SIMILARITIES IN DRUG RESPONSES

This chapter presents a time series data alignment algorithm to identify mechanistic similarities in drug responses, which can facilitate the characterization the mechanisms of action of cancer drugs. We first introduce the background information on the Green Fluorescent Protein (GFP) technology and its usage toward tracking drug responses over time. Then, the rationale of drug comparisons is discussed. Following that, we develop the alignment algorithm in detail. Finally, its performance is tested on a series of real drug experiments.

A. Using Green Fluorescent Protein Technology to Track Drug Responses

The ability to measure the abundance and degree of modification of many macromolecules in cells has allowed researchers to examine cells for particular, existing molecular characteristics that indicate susceptibility to particular drugs [60–62]. This approach has produced a number of very useful guidelines to the use of therapeutics; however, it has failed in instances where the drug induces changes in the type and/or abundance of proteins that either pump drugs out of the cell or proteins that allow the drug-targeted activity to be provided in an alternative way that is not affected by the drug [63, 64]. For these reasons, the ability to examine the molecular dynamics of cells' responses to drugs becomes of primary interest. In addition, identifying those dynamic patterns will help to detect if drugs targeting a particular gene produce or not the desired response. One possible way to achieve this goal would be to develop a way to quantitate the degree of similarity between the responses that cells show when exposed to drugs, so that consistencies in the regulation of cellular response processes that produce success or failure can be more readily identified.

1. Analysis of Gene Transcription Dynamics

A considerable amount of research using fluorescent proteins as transcription activity reporters has examined transcription in living cells in both single-cell and multicellular organisms [65, 66]. Since fluorescent imaging can be carried out in ways that do not destroy cells, fluorescent reporters are very effective tools when studying the time evolution of gene expression. Inserting DNA cassettes with a particular promoter driving expression of a fluorescent protein into egg cells or partially differentiated intermediate cells in ways that allow the cassette to become incorporated in the cell's genome allows one to follow that cell and its daughter cells' developmental course. This kind of information can be used to specify in which cell types and in how many cells a specific gene is active throughout the stepwise course of development and provide clues to gene function. By using a very similar approach on cells with reporters responding to drugs, it is possible to determine which and how many cells are altering the transcription level of a given gene during the course of the cell population's response to the drug. As cells' responses to drugs can take days to run their course, it is expected that a drug that mobilizes a change in the transcriptional regulation of a gene or genes in a cell will produce a distinguishable temporal trajectory of change in both the level of transcriptional change in cells and the number of cells showing altered expression level in a population of treated cells. It is further expected that sets of drugs that induce the same or a very similar alteration in transcriptional regulation will produce similar temporal trajectories, allowing them to be identified as having similarities in their mechanisms of action (MOA).

In our adaptation of this methodology, imaging is carried out using a robotic imaging device (ImageXpress$^{\text{MICRO}}$, Molecular Devices). Multichannel imaging of sets of adherent cells with various reporters cultured in 384 well plates can be carried

out every hour, and at typical initial cell loading levels these cultures can be followed for 50 hours. Two typical fluorescent images are shown in Figs. 29(a) and (b), where nuclei are detected in the blue channel and promoter reporters are detected in the green channel. The objective of image processing is to extract gene-expression levels from the population of cells in the fluorescent image and follow these intensity distributions over time. To do so, we utilize morphological image processing, in particular, the watershed transformation [67]. While the image processing is rather involved, overall it breaks down into three major components: nuclei channel segmentation, reporter channel segmentation, and measurement of cell-by-cell promoter activity levels. Briefly, to extract the promoter activity level for each cell, one needs to first identify the position of cells in the image and then identify the area of the image covered by each cell's cytoplasm. As the cells can be either compact or spread out, we estimate the promoter activity by measuring the sum of the green fluorescent protein (GFP) intensity (arbitrary camera intensity units) in all the pixels within the cell area and reporting the log2 transformed, summed intensity values for each cell. To achieve this, we first process the nuclei (blue) channel to locate all nuclei present in the image, and then process the reporter (green) channel to determine the activity level of the reporter for each cell. We refer to [19] for full imaging details.

Live cell imaging analysis provides two distinct types of cell-by-cell information that are not easily measured over long time spans by other means: (1) the extent of change in promoter activity in the treated population relative to that of the untreated, control population, and (2) the percentage of cells in the treated population shifted into a position in the expression level distribution not occupied by the untreated control population, as a consequence of drug activity. An example of how these two measurements are calculated is presented in Fig. 29(c), where $g(x)$ (control or pre-drug case) and $f(x)$ (post-drug case) represent the log2 GFP intensity distributions

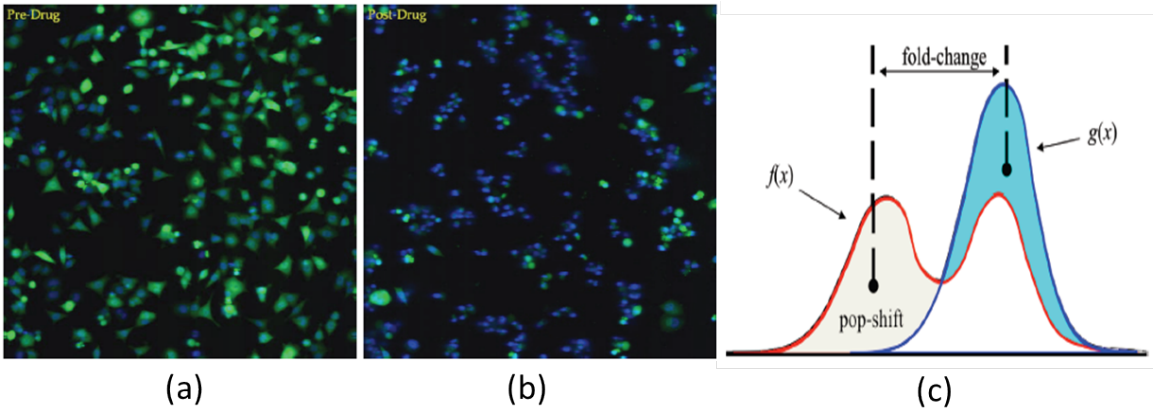


Fig. 29. Two typical fluorescent images for cell-line HCT116 with a promoter reporter for the gene MKI67: (a) before any drug is applied (control or pre-drug case); (b) 43 hours after the drug Lapatinib was added (post-drug case); (c) calculation of population shift/change and fold change: $g(x)$ and $f(x)$ represent the log2 GFP intensity distributions for the cells in (a) and (b), respectively.

for the cells in Fig. 29(a) and (b), respectively. The percentage of population change can be calculated by the difference in area between $g(x)$ and $f(x)$ (the gray area in Fig. 29(c)). Similarly, the fold change can be calculated as the mean difference between the shifted cells and the control case. Note that when fold change is positive, the corresponding population change will be denoted as positive, and when fold change is negative, the corresponding population change will be denoted as negative.

2. What Information on Mechanistic Similarity Is Available in Drug Response Trajectories?

In order to produce metrics of comparison for the similarity of transcription responses induced by drugs, a model of how cellular responses to a drug will shape the trajectories is required. A conceptual model (Fig. 30) of drug response by transcription reporters in one molecularly homogeneous cell line responding to a series of drugs

that target protein regulators acting on pathways of interest facilitates a simplified consideration of the informational content of a trajectory. In this example we start with a set of four genes (A, B, C, and D) that are suspected of contributing to the regulation of a cellular process that we wish to inhibit. We know that a set of genes (E, F, and G) are all strongly expressed when this process is operational and that suppression of expression of gene G produces a reduction in cell proliferation, a desired result of intervention, in these cells. We also have a series of drug compounds, 1, 2, 3, and 4, known to interfere in activation of the transcription factors B, A, D, and C, respectively. In such a setting, Fig. 30(a), the important question to address is if an examination of the dynamics of response to each drug by promoter reporters for genes E, F, and G would produce sufficient understanding of the process mechanisms to determine which genes are driven by a similar regulatory mechanism? As changes in the number of cells making a regulatory decision that leads to altered expression levels allows fairly intuitive interpretation of the dynamics, we will examine this aspect of the conceptual model to illustrate the mechanistic characteristics of the applied drugs that can be inferred through this approach.

If a technical replicate had been run, examining the effect of drug 2 on the gene F reporter, we would expect the kind of trajectories labeled D2 and D2$'$ in Fig. 30(b). (Levels of similarity for replicates in actual experiments are shown in figure on page 77.) If the cell line that these drugs are being tested on is molecularly homogeneous, repeated testing should produce very similar timings of when the cell line will show detectable amount of population change, how rapidly the population change trajectory increases, and how many cells respond to the drug eventually. These three characteristics, time of onset of detectable transcription alteration, rate of population change increase, and final percentage of responded cells define a dynamic population response signature that can be systematically compared across a variety of drugs.

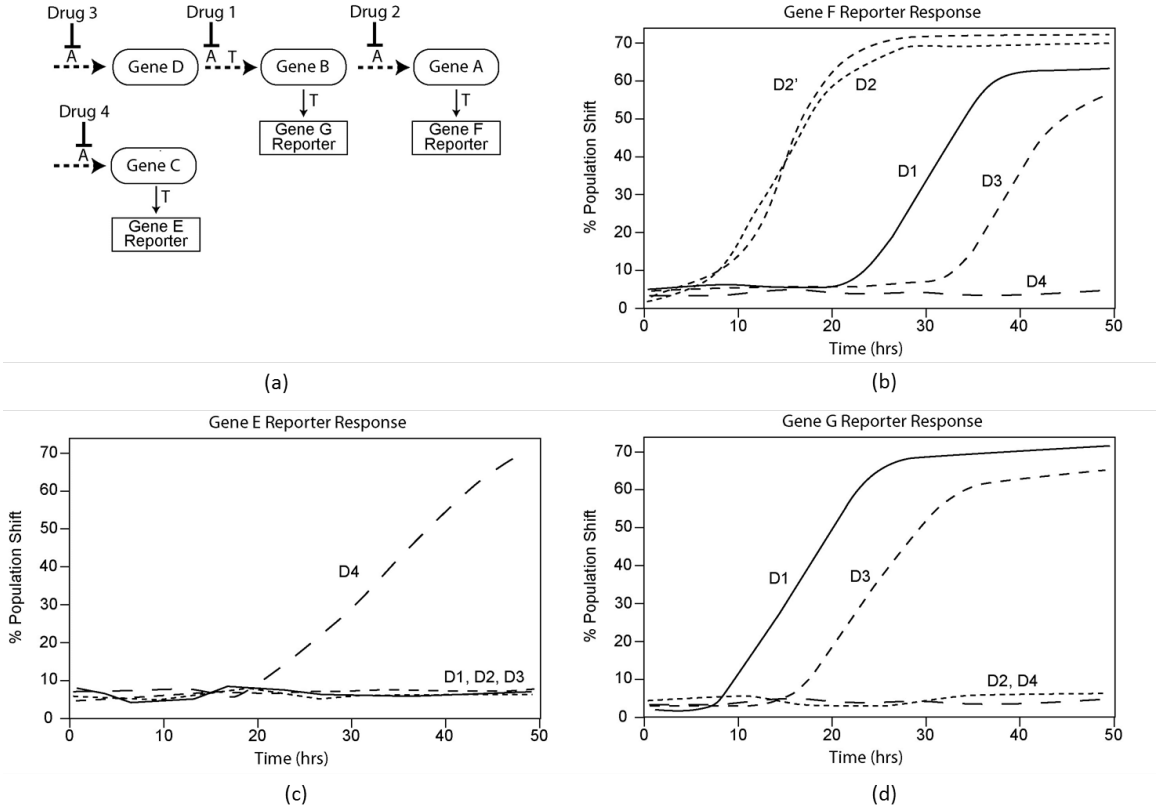


Fig. 30. A variety of possible population change trajectories resulting from drug responses by the pathways diagrammed at the upper left side of the figure. Activation (A) and transcription (T) steps for which components are not shown in this graph occur in the dashed connections between the transcription factors.

If two processes have no overlapping use of components, then the effects of drugs targeting one of the processes should not produce a response from members of the other process. The expected result for this situation is shown in Fig. 30(c). Drug 4 affects process 2 (gene C), but not process 1 (genes D, B and A), so drugs acting on process 1 produce no effects on the process 2 reporter. Similarly, Fig. 30(d) shows that all the three drugs acting on process 1 produce changes in the furthest downstream reporter for process 1 and have no effect on the process 2 reporter.

When drugs are acting on different parts of the same process, it is possible that

the dynamic signatures may be similar. The level of similarity may also vary between reporters placed at different locations along the process, due to differences in both the time required and the step efficiencies in carrying out intervening activation/inactivation, transcription, and translation processes. Differences due to the intrinsic properties of the drugs, rates of cellular uptake, efflux and enzymatic transformation to an active form, where required, could also alter the dynamics of cell response. In Fig. 30(b), responses are shown that could be seen if: (1) drug 2's action on gene B through an inactivation of a transcription factor, gene A, leads directly to shutting down production of the reporter; (2) drug 1's inactivation of transcription factor gene B adds only a single additional transcription step to achieve down-regulation of gene A; and (3) drug 3's rate of inactivation of transcription factor gene D is lower than that of drug 1 and 2 on genes A and B, but the inactivation is very rapidly transmitted from gene D to gene B once achieved. These hypothetical situations and results illustrate the general approaches that could be used to evaluate how the key characteristics of population change relate to cellular drug response dynamics.

B. Rationale for Drug Response Comparisons

To understand the mechanism of action (MOA) of a new drug, it is important to compare its responses to a range of known drugs to see how similar they are. For example, we often have standard drugs that can attack cell lines at various pathways, e.g., survival pathways, proliferation pathways and apoptosis pathways, etc. Understandably, these various drugs will induce different response characteristics of the cell line, which will then be reflected by the different GFP reporters. By comparing to which existing drug the new drug has the most similar response, we can narrow down

the possible MOAs of the new drug. Therefore, there is a need to design proper algorithms that can match mechanistically similar regions for two responses. To design such an algorithm, we first need to understand the nature of drug responses and the types of biological information carried by the various response curves.

A population change responses curve usually has 3 different phases. First is the dormancy period, where the population change is relatively constant and small. In this period, the cell lines are usually making the necessary conditions ready for the target GFP to show any effect. For example, if gene A regulates gene B, the GFP response of B will not change until enough proteins of A have been made. Next comes the responding period, where the right conditions have been prepared and reporters start to be affected by the drug intervention. In this period, we usually see an increase of the population change level. The speed of change depends on a number of factors, including drug dosage. Finally comes the stabilizing period, where the drug has reached its potential and population change is slowed down, eventually reaching a certain level. The final proportion of changed cells depends on the overall efficacy of the drug. An example of the 3 different periods is shown in Fig. 31(a). The reporters start to actively respond to the treatment after ~ 15 hours and reach a plateau after ~ 32 hours. Note that the three steps described here do not necessarily happen for every GFP reporter. For example, if the test drug is ineffective, then the treated cell line will remain in the dormancy period without any significant population change.

The most informative region on any drug response curve is carried in the responding period described above, which is directly affected by a drug's sensitivity and efficacy. Here, we define the *core response* to be the part where population change has attained a certain level (e.g. $> 7.25\%$ population change). We believe that it is meaningful to study a drug's MOA only if it is able to induce a sufficiently population change at some time point during the whole experiment.

To conceptually pose the drug response alignment problem, let us consider the various cases shown in Fig. 31. In the simplest case shown by Fig. 31(a), if two drugs have exactly the same chemical properties, then the response curves produced by the same GFP reporter should look almost identical, as if they were technical replicates. In this case, a direct alignment that matches the same time points on the two responses together should suffice to capture the similarities between them [Fig. 31(a)]. However, this is rarely the case in reality: (1) the slightly different concentrations of the drugs or cell line conditions may cause the two identical drugs to have different lengths of dormancy periods and therefore the responses will have some delays [Fig. 31(b)]; (2) if the two drugs act on the same functional pathway, the one hitting an upstream gene may show earlier responses that the one hitting a downstream gene, again leading to different delays. In such cases, an alignment that allows proper delays should be able to capture the overall similarity level between the two responses [Fig. 31(b)].

A more interesting case is shown in Fig. 31(c), where the two responses show different speeds in increase as well as the final percentage of recruited population percentage. As explained, such cases may be caused by the difference in drug sensitivity and efficacy. The larger difference between the drugs, the quicker the responses deviate from each other. As shown in Fig. 31(c), the alignment will no longer continue after ~ 30 hours.

Another possible pair of responses is illustrated in Fig. 31(d), where the two responses start similar and end similar, but they deviate from each other in the middle portion for a short amount of time. Such cases may be caused by noisy measurements, which break a longer contiguous alignment into two or more smaller pieces.

Real experiments are more complicated and can be a mixture of the various cases

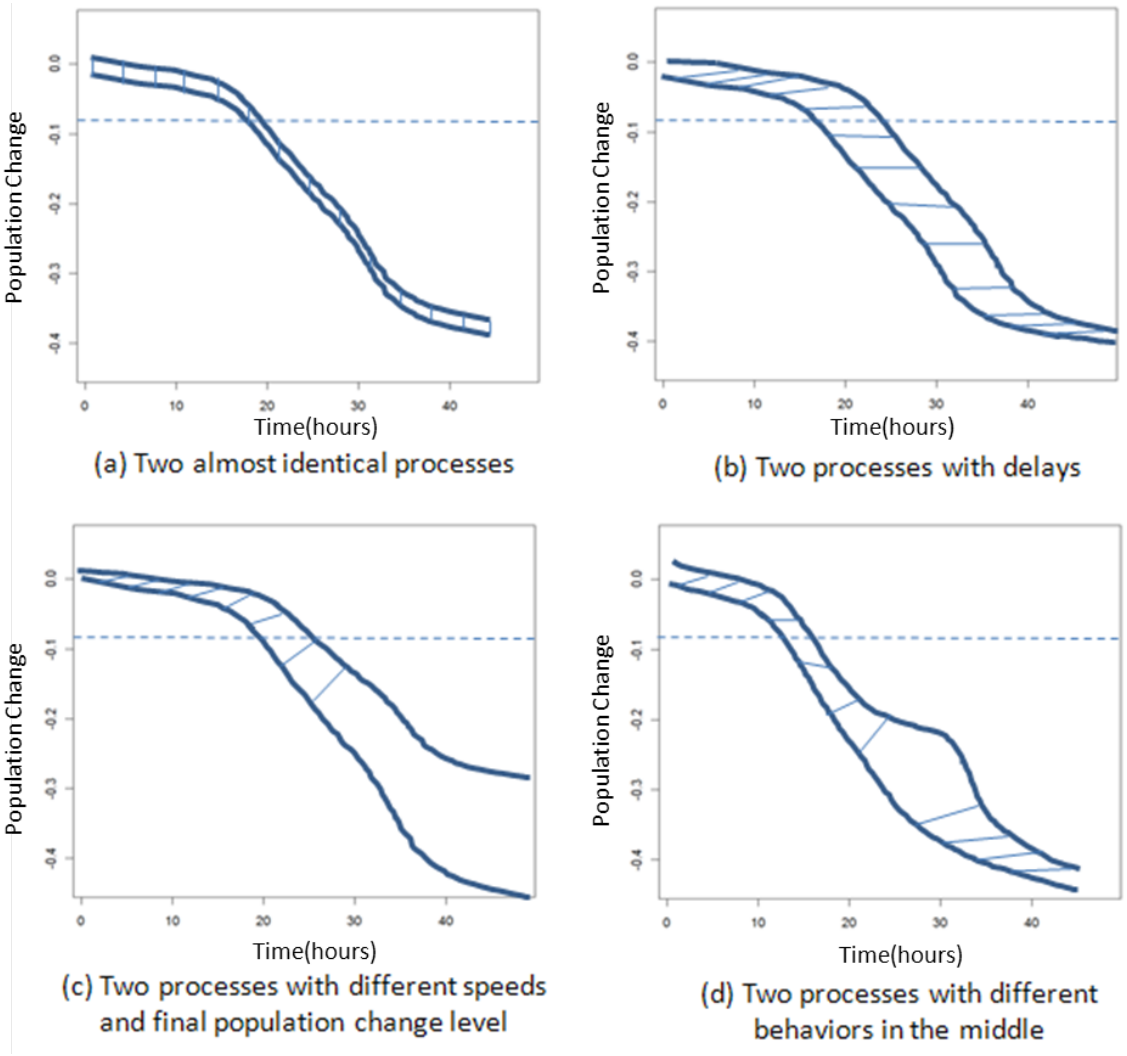


Fig. 31. Conceptualized GFP responses on the population level. (a) Two almost identical responses. (b) Two responses with delays. (c) Two responses with different speeds and final population change levels. (d) Two similar responses with a small portion of difference in the middle. The dashed lines indicate the hypothetical threshold above which the responses can be considered as the core.

described above. Consider the drug experiment shown in Fig. 32, where cell line HCT116 is subject to 3 different drug treatments: AG825, Lapatinib, and LY294002. From a previous study [19], we know that both Lapatinib and LY294002 can attack the autocrine loop involved with the ERBB2/3 heterodimer (higher efficacy for Lapatinib), thereby having very similar MOAs. On the other hand, AG825 mainly works on ERBB2 alone and is not able to break the autocrine loop, thereby having a very distinct MOA from the other two. Looking at the responses for TGFB1 in Fig. 32, Lapatinib and LY294002 show quite similar behaviors on the population change dimension [Fig. 32(a)], with Lapatinib showing stronger responses with a faster speed in the population change. Furthermore, on the fold change dimension, both Lapatinib and LY294002 show very similar overall shape/transition, with a small time delay between the two, while AG825 shows some early similarity but with distinct response after ~ 10 hours.

The following descriptions summarize the key points that determine the mechanistic similarity between two drug responses:

1. The most informative region on a drug response curve is contained in the core region where enough population change has been recruited. The onset time for the core region to happen may vary from drug to drug, however, once they have started, they should proceed side by side. The longer time they continue, the better mechanistic similarity between the two. Therefore, an important criterion to determine the similarity between two drug responses is whether there exists contiguous parts of signals that can be shared by the two responses in their core regions.

2. The core region alignment may naturally extend to regions with lower population change. In Fig. 31 (c), we see that the two drugs are quite similar until

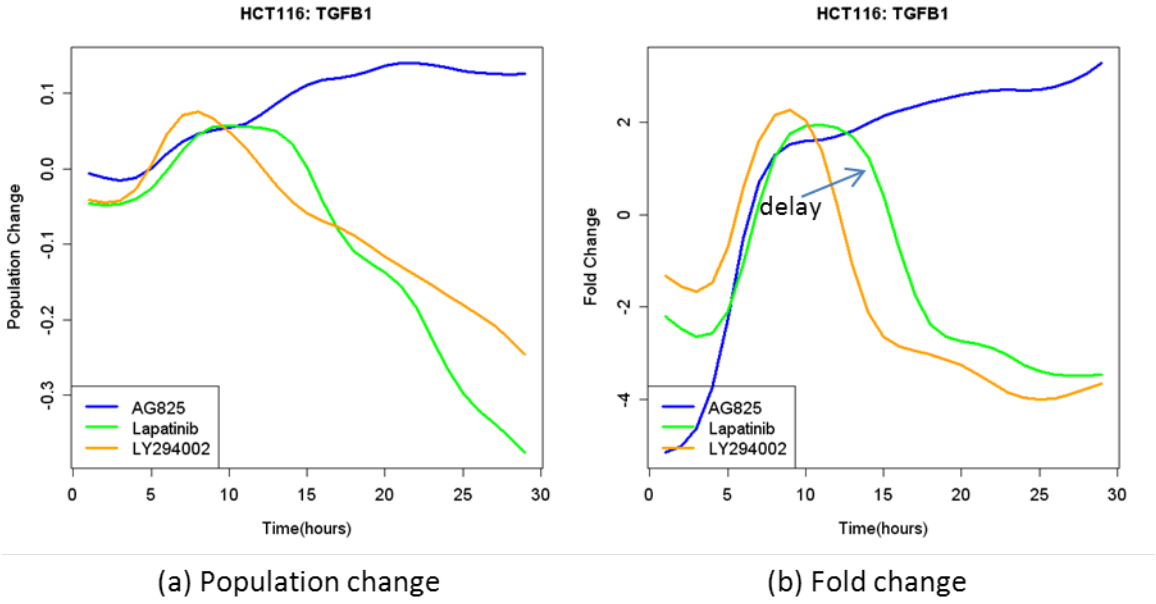


Fig. 32. TGFB1 responses to 3 different drugs on cell line HCT116. (a): population change, (b): fold change. The curves are smoothed by a spline function with 10 degrees of freedom.

$\sim$ 30 hours. Whereas the core regions (defined by the 7.25% population change) only have two pairs of points to be aligned, the total contiguous alignment really starts from the very beginning. Therefore, it is more meaningful to consider the longest contiguous alignment that contains the core region alignment. Here, we define the *core containing alignment* to be the longest contiguous alignment that contains at least one pair of points in the core region.

3. Noisy measurements can break an originally longer contiguous alignment into several smaller pieces. Hence, we should allow small gaps around the core containing alignment region described in point 2. In other words, the similarity comparison for two drug responses should start from the core containing alignment and iteratively search its adjacent regions earlier or later in time, with a small gap allowed (e.g. 2 hours) to account for noise. In the end, the different sections should be aggregated together to reflect overall similarity level. Fig. 31(d) shows an example of such cases.

4. Due to the 2-dimensional nature of the drug response data (population change and fold change), similarity requires the responses to be close on both dimensions. However, in reality, we often observe that when population change is small, the variation of fold change is quite large. Since the calculation of fold change is based on the average behavior of shifted populations, the results might be unreliable when there is only a small population shift. From a previous study, we know that fold change difference can be confidently detected (p-value = 0.05) when the population change is more than 7.25% [19]. Hence, we will only consider the constraint on fold change dimension when the population change has reached 7.25%.

To capture the similarities between reporter responses, we need an "intelligent"

algorithm that is robust to noise, robust to time delays, and finally is able to find all the contiguous parts of signals centered about the core mechanism. Directly comparing points on the two signals is unrealistic, since it cannot handle delays and speed variations on the time axis, as shown in Fig. 33(a). A popular set of algorithms for time series alignment is called the Dynamic Time Warping (DTW) method. It is based on the concept that the similarity between two time series should be computed by locally deforming the time axis in order to minimize the cumulative difference between the aligned points. The DTW algorithm was originally introduced by Sakoe and Chiba for spoken word recognition [22] and it was applied to many other fields including time series gene expression data comparisons [23]. However, there are several disadvantages for the DTW type algorithms. First, for global DTW, all the points on one signal must be mapped to points on the other signal. Thus, outliers cannot be skipped and they can severely distort the alignment. Even though efforts have been made to relax the global alignment constraint, e.g., the open-end DTW algorithms [68], where the head or the tail sections can be left unaligned, are incapable of skipping middle portion outliers if present in the signal. Second, DTW algorithms tend to have many-to-one mappings for the alignment. Therefore, when the two signals are different in amplitude, it is often the case that a large portion of one signal will be mapped to a single point on the other signal to minimize the overall cumulative distance between the two. A global DTW alignment is shown in Fig. 33(b), as shown in the black box, 6 points on the green curve are mapped to the same point on the orange curve, making the alignment very counterintuitive.

A better solution is shown in Fig. 33(c), where the two signals are aligned based on the concept of *Longest Common Substring* (LCSS), which belongs to the class of edit distance problems [69, 70]. LCSS finds the longest string that is a substring of two or more strings. The concept can be extended to real-valued signals in a recursive

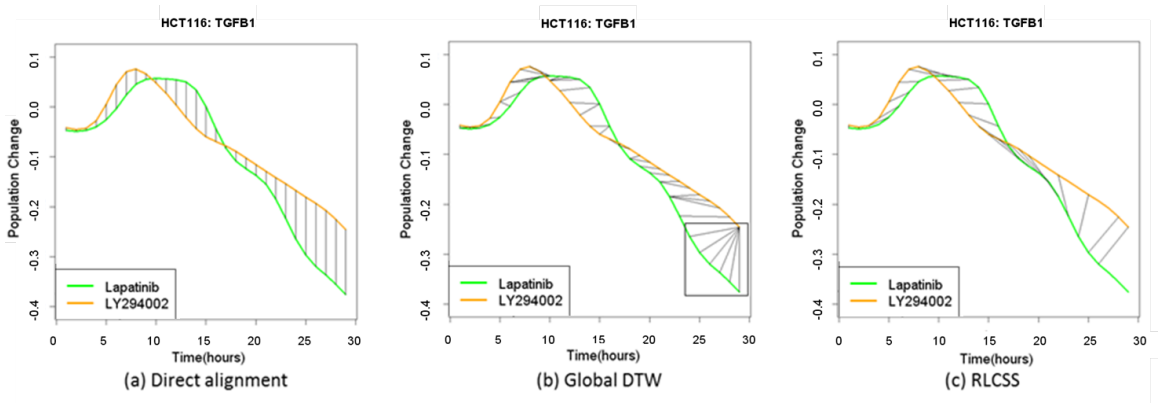


Fig. 33. Direct alignment. It completely ignores the variation in time axis. (b) Global DTW alignment. Many superfluous and spurious matches are seen at the ending sections. (c) RLCSS algorithm. Only one-to-one mapping is allowed and small gaps are allowed to account for noisy measurement.

fashion (the RLCSS algorithm), as we will explain in latter sections. The benefit of RLCSS is that the aligned signals are contiguous and only one-to-one mapping is allowed, which satisfies our assumptions of biological similarity. Furthermore, small gaps (2 hours) are allowed between different sections to account for noisy measurements.

C. Recursive Longest Common Substring Algorithm

1. Definition of Longest Common Substring (LCSS) on Time Series

The original LCSS model refers to a 1D sequence with discrete values, i.e., strings. For example, the sequences ABAB and BABB have their LCSS to be BAB. Our data is 2-dimensional and real-valued. The first dimension is population change and the second is fold change. These reflect two important aspects of the same cell population over time. Therefore, it is natural to consider both dimensions simultaneously when defining similarities.

Formally, let $A = ((a_{x,1}, a_{y,1}), \ldots, (a_{x,m}, a_{y,m}))$ and $B = ((b_{x,1}, b_{y,1}), \ldots, (b_{x,n}, b_{y,n}))$ be two drug responses, where x is the dimension for population change, y is the dimension for fold change, m is the length of A, and n is the length of B. Let $A[1, \ldots, i] = ((a_{x,1}, a_{y,1}), \ldots, (a_{x,i}, a_{y,i}))$.

Definition 1. *Given an integer δ, a real value $k \in [0, 1]$ and a pair of nonnegative real values $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2)$, we define the length of the longest common substring (LCSS) between the two dimensional time series A and B to be the largest element in matrix $R_{\delta,\boldsymbol{\varepsilon}}$, where the element $R_{\delta,\boldsymbol{\varepsilon}}[i, j]$ is defined by:*

$$R_{\delta,\boldsymbol{\varepsilon}}[i,j] = \begin{cases} 0 & \text{if } i = 0 \text{ or } j = 0 \\ 1 + R_{\delta,\boldsymbol{\varepsilon}}[i-1, j-1] & \text{if } |a_{x,i} - b_{x,j}| \le \varepsilon_1,\ |a_{y,i} - b_{y,j}| \le \varepsilon_2,\ |i-j| \le \delta, \\ & |a_{x,i}| \ge k \text{ and } |b_{x,j}| \ge k, \\ & \text{or if } |a_{x,i} - b_{x,j}| \le \varepsilon_1,\ |i-j| \le \delta, \\ & \text{and } |a_{x,i}| \le k \text{ or } |b_{x,j}| \le k, \\ 0 & \text{otherwise} \end{cases}$$

where the constant δ controls the flexibility of matching in time and the constant vector $\boldsymbol{\varepsilon}$ controls the matching threshold and k determines the population change threshold above which fold change constraint will be applied (i.e., the population threshold for the core mechanism). Throughout the paper, we will set $k = 0.0725$, because fold change difference can be confidently detected (p-value = 0.05) when population change has reached 7.25% [19]. Intrinsic to Definition 1 is that $R_{\delta,\boldsymbol{\varepsilon}}[i, j]$ depends only on the previous diagonal element and the current element-wise distance. Hence, $R_{\delta,\boldsymbol{\varepsilon}}[m, n]$ can be efficiently found by filling the tabular starting from $R_{\delta,\boldsymbol{\varepsilon}}[0, 0]$. After the table has been filled, the actual common substring can be found by going back diagonal-wise from the largest entry in the table until a 0 entry is reached

Table IX. The dynamic programming table for finding the LCSS of two 1-D sequences $[0.2, 0.1, 0.1, 0.2]$ and $[0.1, 0.2, 0.1, 0.1]$, with the parameters set to be: $\delta = 4, k = 0, \boldsymbol{\varepsilon} = \varepsilon_1 = 0$. The trace-back path is highlighted in bold face and it contains time indices: $\{(1,2),(2,3),(3,4)\}$

		0.1	0.2	0.1	0.1
	0	0	0	0	0
0.2	0	0	**1**	0	1
0.1	0	1	0	**2**	0
0.1	0	0	2	0	**3**
0.2	0	0	1	0	1

(the trace-back path shown in Table IX). Intuitively, Definition 1 says that 2 drug responses are similar if either of the two conditions satisfies: (1) if population change is large enough, both dimensions have to be similar simultaneously; (2) if population change is not large enough, only population change has to be similar, because the measurement on fold change is no longer reliable. The requirement is made to be consistent with condition 4 described in Section B of this chapter. For illustrative purposes, a numerical example for the LCSS is shown in Table IX. The same operation can be readily extended to real valued sequences as illustrated in Definition 1.

By definition, the substring must be contiguous. This differs from the concept of *longest common subsequence*, where the subsequence is not necessarily contiguous (for example, the longest common subsequence between ABACD and BABD is BAD or ABD). Furthermore, different choices of δ and $\boldsymbol{\varepsilon}$ will lead to different alignment results. A small δ will restrict the alignment points to be close in time. In fact, $\delta = 0$ degenerates to the calculation of direct matching [Fig. 33(a)]. For $\boldsymbol{\varepsilon}$, a very small

threshold will lead to almost no alignment between the two signals, whereas a very loose threshold will lead to every pair of points being aligned as similar. Therefore, an appropriate choice of δ and $\boldsymbol{\varepsilon}$ is important for the application of LCSS alignments. A good choice of the two values depends on the application. Thus, we will have a detailed discussion in the results section.

2. Recursive LCSS Algorithm (RLCSS)

The LCSS algorithm described in the previous section is not yet sufficient for drug response comparisons. First, there is no guarantee that the longest common substring will intersect with the region where sufficient population change has been reached and, therefore, cannot be called the core mechanism described in Section B. Second, even if the core mechanism is found, small gaps should be allowed around it to compensate for noisy measurements. For these two reasons, we define an algorithm that will utilize the LCSS concept recursively to identify the core containing alignment as well as its surrounding pieces. (Note that the trace-back path contains the matched time indices from the two sequences and, therefore, the actual matched points on the two sequences can be directly read out from the path).

1. For a pair of two dimensional sequences A and B, fill the dynamic programming (DP) table as described in Definition 1. Find the longest trace-back path (ending with the largest element in the DP table). If its corresponding matched points contain any member that has sufficient population change on both sequences ($\geq 7.25\%$), then record the trace-back path; otherwise, keep searching the second longest trace-back path in the DP table until it satisfies the population change threshold requirement. Denote the track-back path to be T, where (p, q) is the pair of starting time indices and (s, t) is the pair of ending indices.

Stop and go to step 2. If no such trace-back path exists, exit the program and return $T = NULL$.

2. For the head section sequences $A[1, \ldots, p-1]$ and $B[1, \ldots, q-1]$ of A and B, respectively, fill the DP table to find the LCSS path of the truncated head section sequences. If the ending indices of the LCSS trace-back path is within the time gap allowed from (p, q), then add the newly found trace-back path to the beginning of T; otherwise, continue to search for the second longest common substring until it meets the time gap constraint. Update (p, q) so that it represents the starting indices of the newly formed T and go to step 3. If no such trace-back path is found, stop and continue to step 4.

3. Repeat step 2 for the remaining head sections of A and B until no trace-back path satisfies the condition described in step 2. Stop and continue to step 4.

4. For the tail section sequences $A[s+1, \ldots, m]$ and $B[t+1, \ldots, n]$ of A and B, respectively, fill the DP table to find the LCSS path of the truncated tail section sequences. If the starting indices of the LCSS trace-back path is within the time gap allowed from (s, t), then add the newly found trace-back path to the end of T; otherwise, continue to search for the second longest common substring until it meets the time gap constraint. Update (s, t) so that it represents the ending indices of the newly formed T, and go to step 5. If no such trace-back path is found, stop and exit program.

5. Repeat step 4 for the remaining tail sections of A and B until no trace-back path satisfies the condition described in step 4. Stop and exit program.

Note that by aligning the head sections and tail sections separately, it is guaranteed that the time order will not be destroyed. Fig. 34 shows a graphic illustration

of the RLCSS algorithm.

Definition 2. *The similarity $S_{\delta,\varepsilon}(A, B)$ expressed in terms of the RLCSS similarity between the time series A and B is given by:*

$$S_{\delta,\varepsilon}(A, B) = \frac{|T|}{min(n, m)}$$

where T is found by the RLCSS algorithm and $|T|$ is the cardinality of T. Note that $S_{\delta,\varepsilon}(A, B)$ is always between 0 and 1 – the larger the value, the greater the similarity.

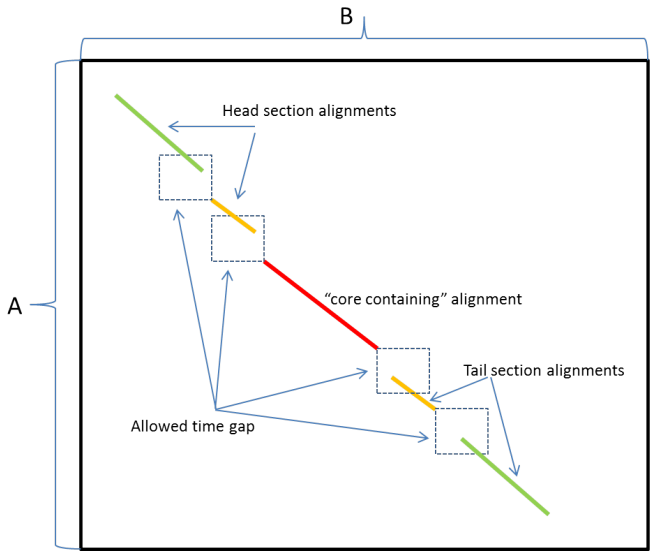


Fig. 34. Illustration of the RLCSS algorithm. The DP table is represented by the big solid black box. The algorithm starts by finding the core containing alignment (red solid path), and subsequently recursively finds the head section alignments and tail section alignments (orange and green paths) around it with small time gap allowed (dashed boxes). In the end, the alignment path will include all the 5 sections.

D. Results for RLCSS Alignment

In this section, we apply the RLCSS algorithm to drug response data. To test its performance, we need to know a priori the MOA of the testing drugs and see whether the alignment results agree with our prior knowledge. For instance, if we know that drug X and Y have very similar effects on some cell line (due to similar MOAs), can the proposed RLCSS algorithm also capture their similarities and claim they are similar? Conversely, if drug X and Z are very different in their MOAs, is the RLCSS algorithm also able to claim that they are dissimilar?

To test the performance of the proposed RLCSS algorithm, we consider the results of a study in which the detailed MOA of each drug has been carried out [19]. In this study, 5 drugs (Lapatinib, LY294002, Temsirolimus, U0126, AG1024) are tested against the cancer cell line HCT116. Referring to Fig. 35 and ranking drug responsiveness across the drugs for HCT116, one observes responses over most reporters but with decreasing percentages of cells shifted for Lapatinib (EGFR/ERBB2), LY29004 (PI3K) and Temsirolimus (mTOR). This similarity of action with decreasing efficacy falls directly along a survival signaling pathway headed by an activated receptor heterodimer, ERBB2/ERBB3, and then proceeds along the canonical PI3K/AKT/mTOR pathway. The remaining drugs' inhibitory powers would be ranked U0126 (MEK1/2) and AG1024 (IGF1R) – AG1024 not shown in the figure because it acts on a kinase not shown in the figure. All of these second tier drugs deliver very low reductions of transcription of MKI67, the current "gold standard" [71, 72] in tumor pathology for determining the proliferative state of a tumor. The results show that Lapatinib, LY294002 and Temsirolimus have similar MOAs in the sense that they all target the same survival pathways. On the other hand, U0126 and AG1024 have very dissimilar MOAs compared to the three previously mentioned drugs.

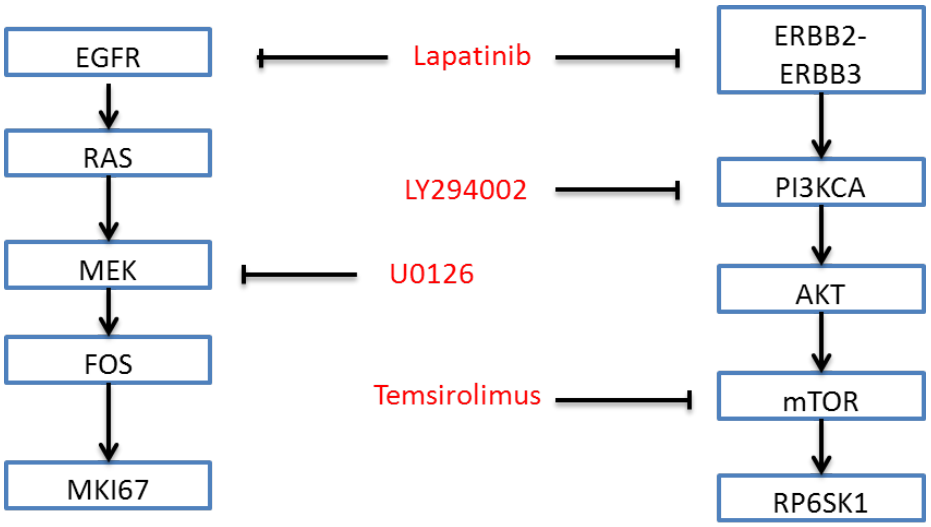


Fig. 35. Relative strengths of drugs versus position in pathways and inferred crosstalk between survival and proliferative signal channels.

We design several sets of experiments to test if the proposed RLCSS algorithm reaches the same conclusions as just described. First, to get a sense of the variation presented in the drug response data, we test RLCSS on a set of technical replicates (TR), with the idea that TRs should exhibit high degrees of similarity among each other. Furthermore, by studying the TRs, we can to find a proper range for the two key parameters in the RLCSS algorithm. Second, we test RLCSS on Lapatinib, LY294002 and Temsirolimus, since they are related in their MOAs, and the degree of similarity should be high, but not as high as with the TRs. Last, we test RLCSS on Lapatinib, LY294002, U0126 and AG825, because the first two drugs are very different from the last two in their MOAs. We set the parameter $k = 0.0725$ in Definition 1 and the time gap to be 2 hours for all experiments.

1. RLCSS Performance on Technical Replicates

For RLCSS to work in practice, a proper set of values must be determined for δ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2)$. δ controls the maximum time delay allowed for the matching. In practice, the experiment lasts around $30 - 50$ hours and we know that transcription is a relatively slow process. It usually takes about $8 - 12$ hours for a reporter to show some activity after the initial treatment. Therefore, in our application, we set δ to be between $8 - 12$ hours, which should be enough to compensate for the time delays of different drugs. Moreover, we observe that RLCSS is usually insensitive to δ variations in the sense that changing δ in that range does not change the alignment results significantly (See Table X). $\boldsymbol{\varepsilon}$ determines the threshold for similarity. The idea for determining $\boldsymbol{\varepsilon}$ is to set it to be large enough to compensate for the discrepancies in technical replicates. Here, we set $\boldsymbol{\varepsilon}$ to be the value so that the worst case technical replicates similarity is at least 75% (Fig. 36, black and red curves). In practice, we have found that it is enough to account for the biological variations with ε_1 close to 0.09 and ε_2 close to 0.8, as we show in Table X.

Figs. 36(a) and (b) show a set of 4 technical replicates on the reporter MKI67 on the cell line HCT116 after treatment with Lapatinib, and Figs. 36(c) and (d) show the RLCSS alignment for the worst case TR similarity. The pairwise alignment result is summarized in Table X. The alignment results are not affected at all by the different choices of δ ranging from 10 hours to 12 hours for all the pairwise alignments except dup1 and 4, indicating that the RLCSS algorithm is very robust to δ variations. As seen in Fig. 36 (d), the RLCSS algorithm successfully identifies the time delays between the two replicates.

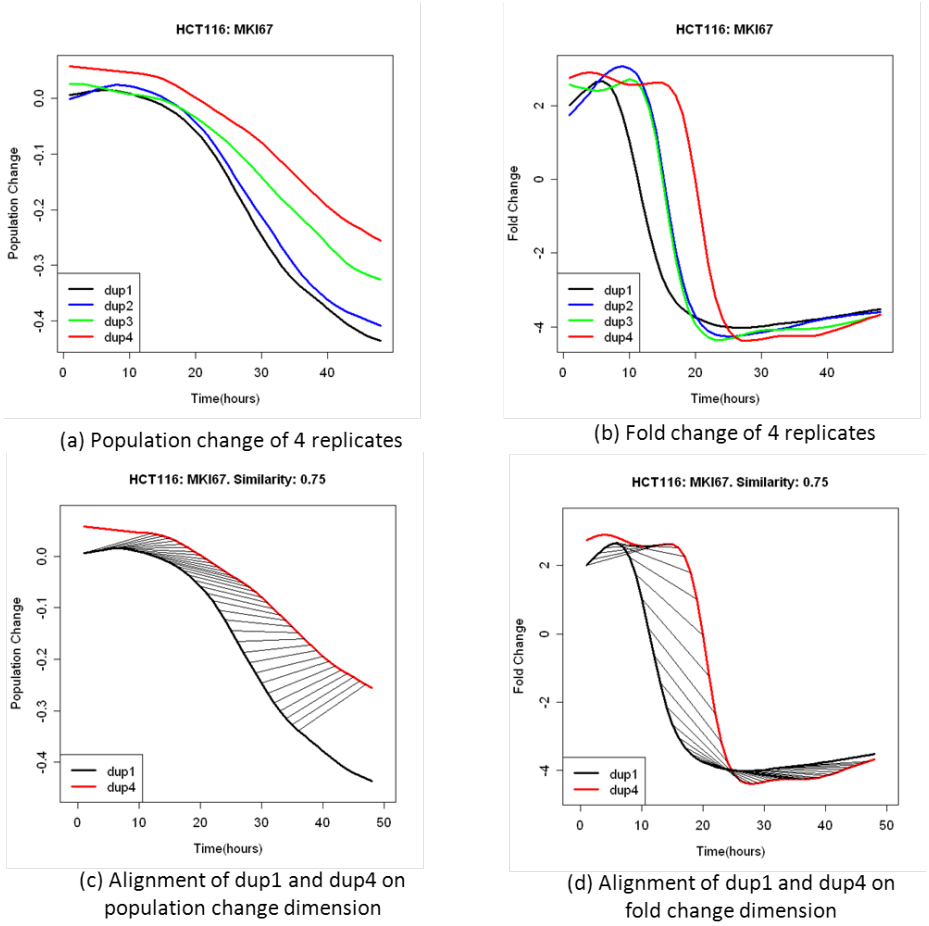


Fig. 36. Technical replicates of Lapatinib treatment on cell line HCT116. ε is set to be the value so that the worst case technical replicates (black and yellow) similarity is at least 75%.

Table X. Pairwise similarity between technical replicates, with different δ. As can be seen, the different choices of δ do not affect the alignment results significantly.

	Dup1, 2	Dup1, 3	Dup1, 4	Dup2, 3	Dup2, 4	Dup3, 4	δ	ε
Similarity	1	0.917	0.667	0.958	0.75	1	8	(0.09,0.8)
Similarity	1	0.917	0.688	0.958	0.771	1	9	(0.09,0.8)
Similarity	1	0.917	0.708	0.958	0.792	1	10	(0.09,0.8)
Similarity	1	0.917	0.75	0.958	0.792	1	11	(0.09,0.8)
Similarity	1	0.917	0.75	0.958	0.792	1	12	(0.09,0.8)

Table XI. Pairwise similarity between 3 drugs with similar MOAs, with $\delta = 11$ and $\varepsilon = (0.09, 0.8)$.

	Lapatinib LY294002	Lapatinib Temsirolimus	LY294002 Temsirolimus
TGFB1	0.862	0	0.138
ERBB3	0.862	0.724	0.793
EGR1	0.611	0.167	1
MKI67	0.621	0.69	0.897
FOS	0.677	0.583	0.25

2. RLCSS Performance on Lapatinib, LY294002 and Temsirolimus

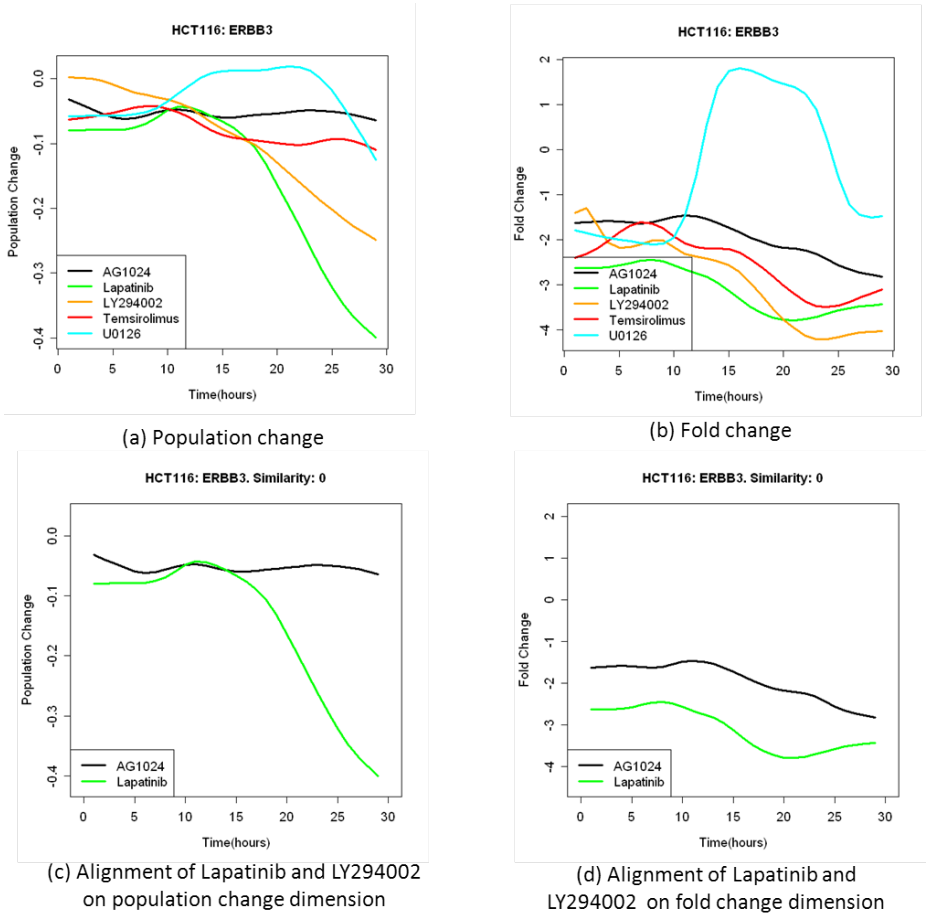
We design the second set of experiment to show the utility of RLCSS on 3 drugs with similar MOAs. As described in Section B, it is meaningless to compare drug responses on reporters whose responses have not changed enough during the whole experimental span. Therefore, we select the reporters whose responses show at least 7.25% population change for at least 2 out of the 3 drugs. The similarity comparison table of the 3 is summarized in Table XI. The 3 drugs show considerable amount of similarity with each other, especially on ERBB3 and MKI67, the two key reporters that reflect the MOA of drugs on cell line HCT116 [19]. We also observe that Lapatinib is closer to LY294002 than it is to Temsirolimus. The result is also consistent with our prior knowledge that LY294002 is closer to Lapatinib than Temsirolimus in their actual positions of attack (Fig. 35).

3. RLCSS Performance on Lapatinib, U0126 and AG1024

The third set of experiments is intended to test whether the RLCSS algorithm is able to detect mechanistic difference between drugs. As we know from earlier discussion, Lapatinib is very different from U0126 or AG1024 in their MOAs. The similarity results of the 3 drugs are summarized in Table XII. Ranking by the closeness to Lapatinib, the order of similarity is: U0126 and AG1024.

4. RLCSS to Detect Apoptosis

It is possible that the reporter responses are different in the beginning, but later in time behave similar due to a common process, e.g. apoptosis. This is because once a cell has determined to go through apoptosis, all the reporters will have a significant drop in their activity level and eventually die out. UNBS1415 is a drug,



(a) Population change

(b) Fold change

(c) Alignment of Lapatinib and LY294002 on population change dimension

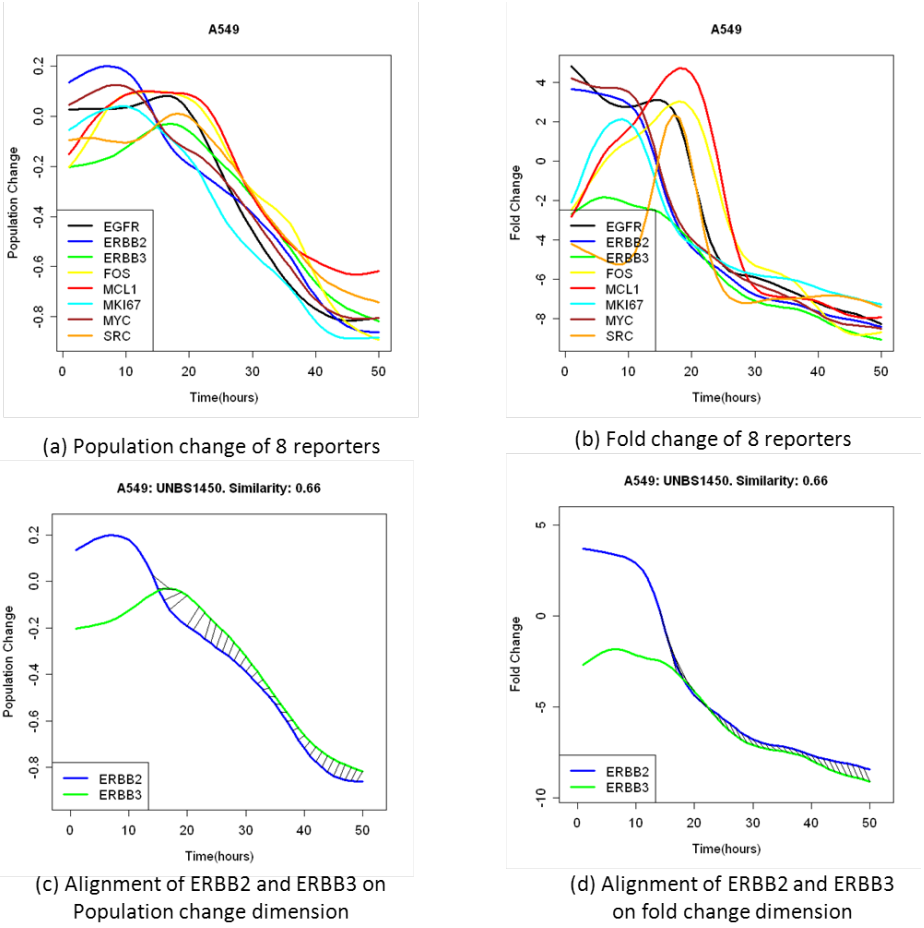
(d) Alignment of Lapatinib and LY294002 on fold change dimension

Fig. 37. Responses of ERBB3 to 5 different drugs. Looking at the figure, it is not surprising to see why Lapatinib has 0 similarities with AG1024. For example, the black curve (AG1024) has a very small population change during the entire experiment and therefore it forms no core mechanism alignment with Lapatinib, even though its early population change is quite close to Lapatinib. The RLCSS algorithm has the advantage to filter out "uninteresting" similarities.

Table XII. Pairwise similarity between 3 drugs with distinct MOAs, with $\delta = 11$ and $\boldsymbol{\varepsilon} = (0.09, 0.8)$.

	Lapatinib U0126	Lapatinib AG1024	U0126 AG1024
TGFB1	0.793	0	0
ERBB3	0	0	0
EGR1	0.306	0	0
MKI67	0	0	0
FOS	0.583	0.611	0.861

that induces apoptosis on the cell line A549. In Figs. 38(a) and (b), we can see that the initial responses are quite different for different reporters; however, later in time, all responses seem to converge to the same behavior after $25 - 30$ hours. In Figs. 38(c) and (d), we see that RLCSS algorithm successfully identifies the similarity later in time, which is exactly what we expect. In fact, RLCSS algorithm is able to detect all the similarities later in time for all pairs of responses in this example. For space considerations, we only show one example here.



(a) Population change of 8 reporters

(b) Fold change of 8 reporters

(c) Alignment of ERBB2 and ERBB3 on Population change dimension

(d) Alignment of ERBB2 and ERBB3 on fold change dimension

Fig. 38. Responses of 8 reporters to the drug UNBS1450 on cell line A549. Note that UNBS1450 is able to induce apoptosis on this particular cell line and therefore, all the later responses are very similar for all the reporters. The RLCSS algorithm successfully identified the similarity later in time. The parameters are $\delta = 11$ and $\boldsymbol{\varepsilon} = (0.09, 0.8)$.

CHAPTER IV

MODELING POPULATION OF CELLS' GENE EXPRESSION DYNAMICS AFTER DRUG TREATMENT

This chapter presents a Markov model to describe a population of cells' behavior after drug treatment. To begin with, we introduce the asynchronous and independent nature of cells' responses to drug treatment. Then, we propose and develop the Markov model in detail. The estimation of model parameters are also discussed. In the end, we explain how to use the model for experimental design.

A. Gene Expression Varies from Cell to Cell

Since gene expression happens within individual cells and the traditional way of microarray experimentation measures average gene expression from a mixture of cells, the results may not be reflective of the true state of gene activities. In fact, many recent research findings note the limitations of a notion such as "average cell" because there is significant variation among different cells – even for the same cell line in a hypothetically constant and homogeneous intracellular environment [73, 74]. One possible contributing factor is the inherent randomness associated with transcription, e.g. chromatin remodeling. In particular, gene expression is governed by "transcription bursts": a gene stays a long time in the inactive state, followed by a short period of activity where it makes a burst of transcripts [74]. Because such bursts happen randomly within different cells, the number of transcripts also varies from cell to cell.

Given the inherent randomness associated with gene expression among different cells, it is preferable to study the gene expression distribution instead of some average value. One apparent benefit is that it can reveal subpopulation differences (if any) to external stimuli (e.g. drugs), which is a key step in understanding cell

dynamical response to drug treatments. Recently, [19] described an automated system that allows researchers to track the transcriptional activities of multiple genes under different external stimuli for extended periods. Briefly, the coding sequence of a Green Fluorescent Protein (GFP) is fused with the promoter region of a gene of interest. Subsequently, a single cassette bearing the promoter/GFP is delivered into the genome of each cell in a population of cells (note that the insertion point of the cassette is highly random among different cells). Any change in the expression levels of the native coding sequence driven by that promoter will be reflected by the intensity of the GFP reporter. The activity of the GFP reporter is captured by digital microscopes at hourly intervals for ~ 50 hours. Two typical snapshots of the GFP images are shown in Fig. 39, parts (a) and (b). In the end, image processing algorithms segment GFP intensity for individual cells and the gene expression for the cell population is reflected by the intensity distributions of the respective GFP reporters (Fig. 39(c)). Interested readers should refer to [19] for a detailed description for the experimental protocols and image processing algorithms.

Two features can be observed from Fig. 39(c). First, the GFP intensity distribution appears bimodal after drug treatment rather than gradually moving to the left as a unimodal distribution. This indicates that each cell make a large shift in its transcription level independently of other cells rather than incrementally decreasing its transcription level in synchronization with other cells. This suggests that a two-state model could be used to describe the underlying gene expression activities. Second, within each state, the actual transcription level varies from cell to cell (otherwise, the gene expression would be two narrow spikes instead of two relatively wide bell shape curves). The randomness could be attributed to the inherent variations caused by "transcription burst" and to the different transcriptional efficiencies resulting from the random insertions of the promoter/GFP cassettes into different cells.

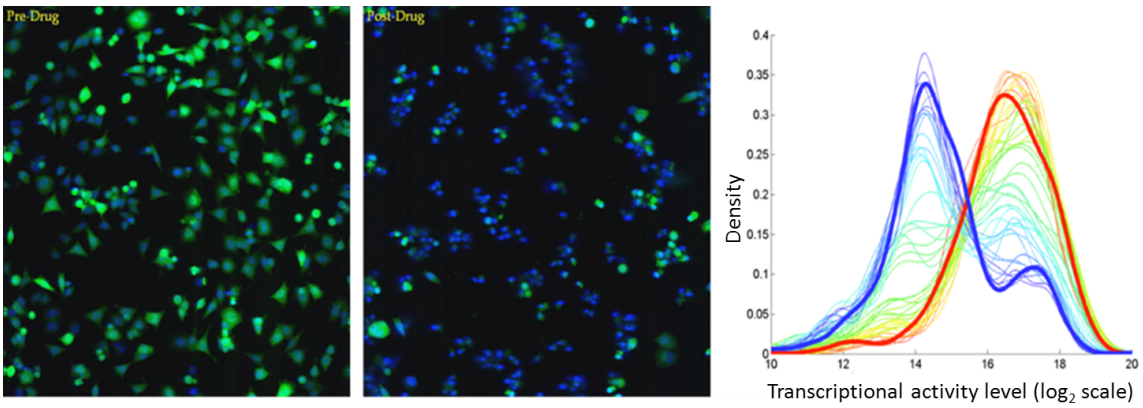


Fig. 39. Fluorescent images of the same imaging site for cell line HCT116 with a promoter reporter for the gene MKI67 taken (a) before any drug was applied, (b) 43 hours after the drug Lapatinib was applied (Green color indicates the activity of GFP reporter and blue indicates the location of nuclei). (c), the GFP log2 intensity distributions for the same cell line at various time points (time is color coded starting from red, changing to yellow and green and finally blue).

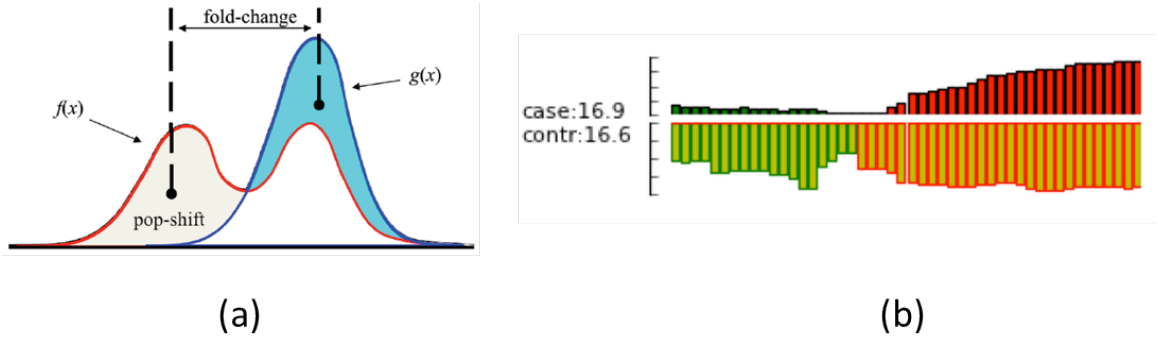


Fig. 40. Measuring the relative transcription activity difference through pop-shift and fold-change. The gray area under the red distribution represents the shifted cell population percentage compared to the blue distribution. The difference between the mean of gray area and blue area is the corresponding fold-change. (b) Bar-plots show the relative transcription activity of the cell population of Fig. 39(c) throughout the experiment with the control population set to the un-drugged population at the same time point. The drug was added after 5th hour. The top bar-plots show the pop-shift, while the bottom ones show the fold-change. Each tick in y-axis of PS plots corresponds to 10% shift, while each tick in fold change corresponds to a 2-fold concentration change from previous tick. The green bars indicate up-regulation while the red bars indicate down-regulation. The expression level of the initial state for both case and control are shown at the left of each plot.

Given the time course gene expression distributions shown in Fig. 39(c), it is possible to compute two types of measurements that describe their dynamics in a concise way: the population-shift (PS) and fold-change (FC) (Fig. 40(a)). PS describes the percentage of shifted cells relative to control (un-drugged case) at any given time, and FC describes the extent of gene expression change of the shifted cells. These two values can be plotted over time to reflect cell changes due to external stimuli (Fig. 40(b)).

B. Modeling the Gene-Expression Distribution of a Cell Population

Mathematical models facilitate systematic understanding of the gene-expression distribution data generated by the GFP technology. A good model should possess the following properties:

1. It should be able to emulate the two-state asynchronous transition of individual cells, while allowing for gene expression uncertainty.
2. It should contain biological relevant parameters essential to the observed process and there should be a procedure describing how the model parameters can be inferred from experimental data.
3. It should be capable of generating useful hypotheses for future experimental design.

1. A Two-State Random Process to Describe Single Cell Behavior

As observed in experiments, a cell makes its decision asynchronously and independently of the other cells. Therefore, in a homogeneous cell line, it suffices to assume that the cells are i.i.d. The population behavior will be determined by the combination of the dynamics of each single cell, just as the number of heads in an experiment of flipping coins from a pool of i.i.d. coins is determined by the individual coin's probability of landing head. We first consider the key parameters that describe the gene-expression distribution for a population of cells:

1. Total number of cells: N.
2. Mean and Variance of the initial intensity distribution (State 1): μ_1, σ_1.
3. Mean and Variance of the post-drug intensity distribution (State 0): μ_0, σ_0.

4. Rate of transition from state 1 to state 0 (or equally, the transition probability): c.

5. Final proportion of responded cells: K.

6. The onset response time for each cell: t_0.

The total number of cells N is determined by the experimental set-up. In a typical GFP study, it is usually $300 - 400$ cells. The means and variances of each state are determined by the inherent randomness described in the introduction section and are usually different from gene to gene. The rate of transition could depend on the concentration/dosage of the drug applied to the cell line. Presumably, cells will transition at a lower rate when lower dosage is used (but the dependency might not be linear). The final proportion of responding cells measures the overall efficacy of that drug and is directly related to the rate of transition. Finally, the onset response time for each cell is defined to be the time needed for the cytoplasmic concentration of a drug to reach a threshold level so that the drug can actually take effect. Note that the exact level of the threshold is not important for our consideration and that the onset time really reflects the fact that, upon adding the drug, it takes certain amount of time for the drug to permeate the cell membrane and accumulate enough so that the cell will actually be affected. The onset time could be different for cells in a homogeneous cell line, because each cell is subject to different micro-environments, such as the nutrients to which it is exposed and the cell cycle state it is in. Therefore, the onset times of different cells reflect the "readiness" of particular cells to be transformed by the drug.

Consider any single cell, for gene i, at time t, its expression value, $r_i(t) \in R$, is a real number and its expression state, $x_i(t) \in \{0, 1\}$, is assumed to be binary (the subsequent theory extending to any discrete state space). Note that the time

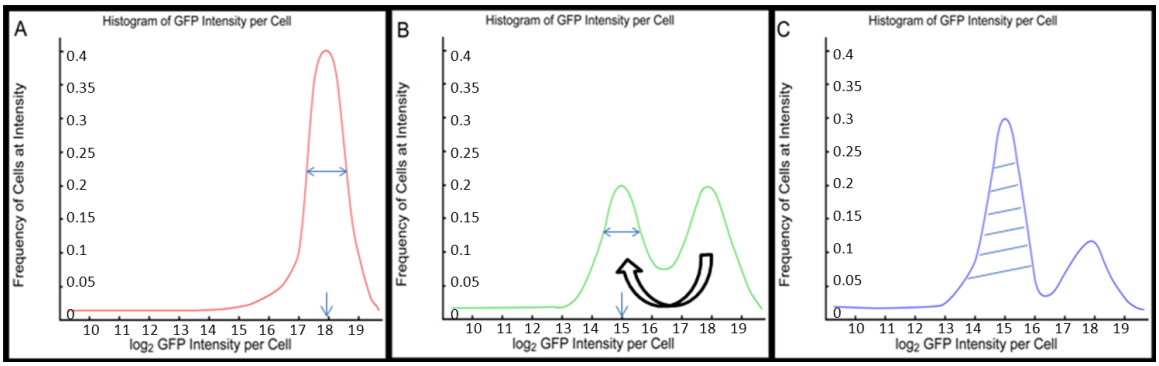


Fig. 41. Schematic figures showing the key parameters that determine the dynamics of gene expression distributions: Means and variances of state 1 and 0, transition rate, and final proportion of responded cells.

t is not continuous in our experiment. It is discretized to reflect the sampling rate when the GFP images are taken by the digital microscope: $t = 0, 1, \ldots, T$, where T is the duration time of the experiment. Whenever it is clear from the context, we omit the time dependency. The expression value is governed by the expression state and assumed to be randomly sampled from a Gaussian distribution whose mean and variance are time invariant and are only determined by the corresponding expression state of that gene i: $r_i(t) \sim N(\mu_{x_i}, \sigma_{x_i})$. Denote $p_0^i = P(x_i = 0)$ and $p_1^i = P(x_i = 1)$ and $\boldsymbol{p}^i = [p_0^i, p_1^i]$.

Drugs act as external inputs for the cell and can break a cell's hemostasis balance by blocking/inducing the expression of the respective target gene(s). To model drug effectiveness, we assume that the drug effect on its target gene i is either ineffective or effective, namely $y_i(t) \in \{0, 1\}$. When $y_i = 0$, the respective target gene remains in its original hemostasis balance state; when $y_i = 1$, that balance is disturbed, which leads to a new transition period, and possibly a new hemostasis state. Similarly, a gene j not directly affected by the drug but affected through signaling cascades will respond to its transcription factors' net effect. In either case, for any gene in the

cell, y_i can be viewed as the net effect of its regulators. Hence, we will call y_i the regulation state of gene i. A gene's regulation state should affect its expression state, as we explain next.

Here, we propose a two-state Markov model to describe the ith gene expression-state dynamics with respect to its regulation state (net effect of drug or transcription factors), y_i. Let A_{i,y_i} be the transition probability matrix of gene i with regulation state y_i:

$$A_{i,y_i} = \begin{bmatrix} a_{00}^{i,y_i} & a_{01}^{i,y_i} \\ a_{10}^{i,y_i} & a_{11}^{i,y_i} \end{bmatrix} \tag{4.1}$$

where a_{mn}^{i,y_i} is the transition probability of gene i transition from the expression state m to expression state n, with regulation state y_i. We have $a_{00}^{i,y_i} + a_{01}^{i,y_i} = 1$ and $a_{10}^{i,y_i} + a_{11}^{i,y_i} = 1$. Fig. 42 illustrates the two-state model proposed. Note that, given y_i, we implicitly assume that A_{i,y_i} is time invariant, however, in a more general case, A_{i,y_i} can also be time dependent.

To update the expression state of gene i, we have:

$$\boldsymbol{p}^i(t+1) = \boldsymbol{p}^i(t)A_{i,y_i(t)=0}1_{y_i(t)=0} + \boldsymbol{p}^i(t)A_{i,y_i(t)=1}1_{y_i(t)=1} \tag{4.2}$$

where $1_{y_i(t)=0}$ or $1_{y_i(t)=1}$ is the indicator function. Eq. (4.2) says that, in any cell, the expression state $x_i(t+1)$ of gene i at time $t+1$ is determined by its previous expression state $x_i(t)$ and regulation state $y_i(t)$, through the corresponding transition probability matrix A_{i,y_i}. Note that the regulation state $y_i(t)$ is not necessarily Markovian and it only reflects the net effect of the regulators of gene i. We subsequently return to the discussion of $y_i(t)$.

The model proposed here describes the experimental observations that, for a gene to change its expression state, all necessary conditions must already be in place,

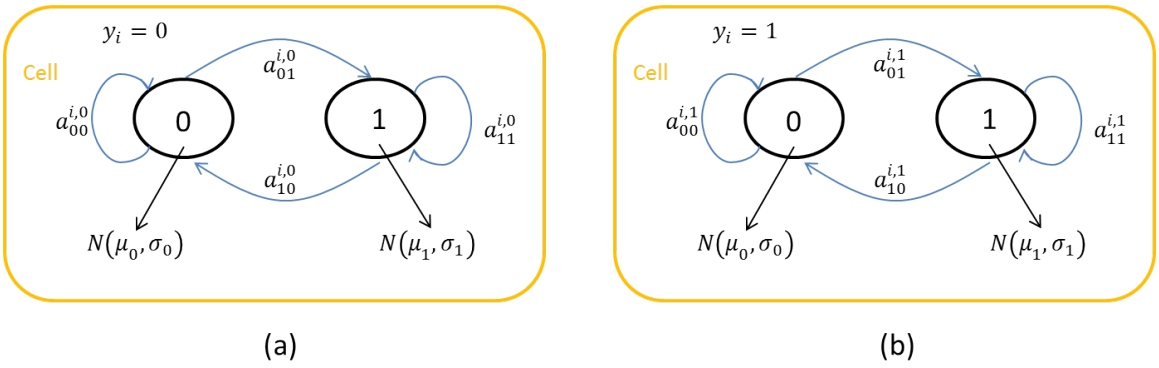


Fig. 42. Two-state Markov model to describe a gene i's gene expression dynamics with respect to its regulation state $y_i = 0$ or 1. Note that the gene expression state transition probabilities depend on the regulation state of that gene.

i.e., $y_i = 1$. Upon such conditions, the gene will switch its expression state probabilistically.

2. Constant Onset Time for Each Cell in the Cell Population

Let us consider the relationship that a drug inhibits the expression of gene i. In the simplest condition, we can assume that the onset time for each cell t_0 is a constant. Furthermore, experiments suggests [19] that prior to adding drugs, the cells in a particular cell line often show a unimodal gene expression distribution (see the thick red curve in Fig. 39 (c)), indicating that all cells are in the same expression state. Therefore, in our model we assume $p_0^i = 0$ and $p_1^i = 1$ to be the initial probability distribution of the respective gene expression state. That is, all cells are concentrated in the high expression state, $x_i(t) = 1$, for any $t \leq t_0$. During that period, there is a negligible likelihood, i.e., 0, probability, for transitioning from the high-expression state to the low-expression state. Hence, $a_{10}^{i,0} = 0$ and $a_{01}^{i,0} = 0$. After the onset time t_0, drugs start to take effect $y_i = 1$, and therefore cells begin to transform from

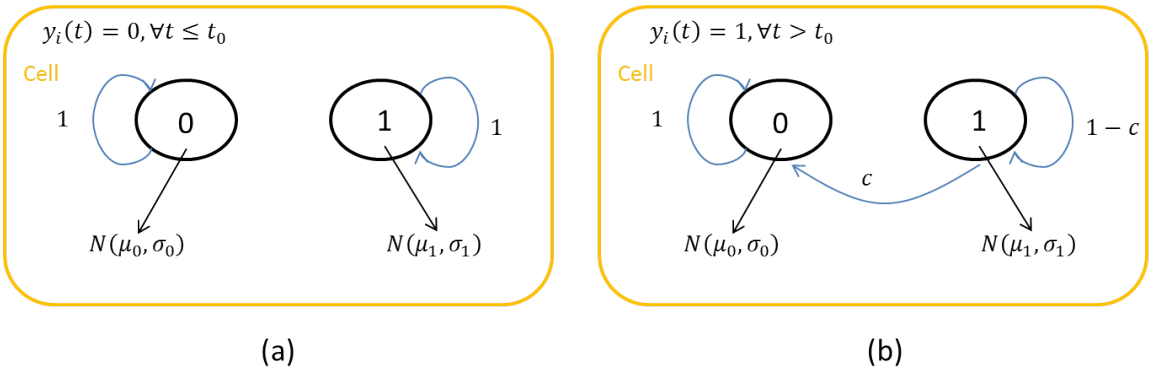


Fig. 43. A simplified two-state transition model, assuming identical onset time t_0 for each cell. (a) Before the onset time, all the cells stay in the high expression state, with no probability of transition to the low expression state. (b) After the onset time, with constant probability c, a cell will transit from the high expression state to the low expression state.

high-expression state to low-expression state, with a constant transition probability, $a_{10}^{i,1} = c$. On the other hand, once a cell has made its transition, it will stay at the low-expression state with null probability of returning back, i.e., $a_{01}^{i,1} = 0$, as suggested by the experimental data. The summary figure of these conditions is illustrated in Fig. 43.

Given this simplified state transition model, it is interesting to compare theoretically derived results with data from experiments. Fig. 44 shows an example of the simulation results generated from the transition model described in Fig. 43, with $N = 400, T = 20, \mu_1 = 15, \sigma_1 = 2, \mu_0 = 8, \sigma_0 = 2, t_0 = 5, c = 0.2$. Looking at Fig. 44, we see that there is no population shift before the onset time $t_0 = 5$, since we have assumed $a_{10}^{i,0} = 0$. However, there is one apparent discrepancy between the simulation results and the experimental results, namely, the number of transformed cells per unit time is highest right after the onset time and gradually decreases as time goes on. Such behavior is not accidental, since at time $t = t_0 = 5$, all the cells are ready

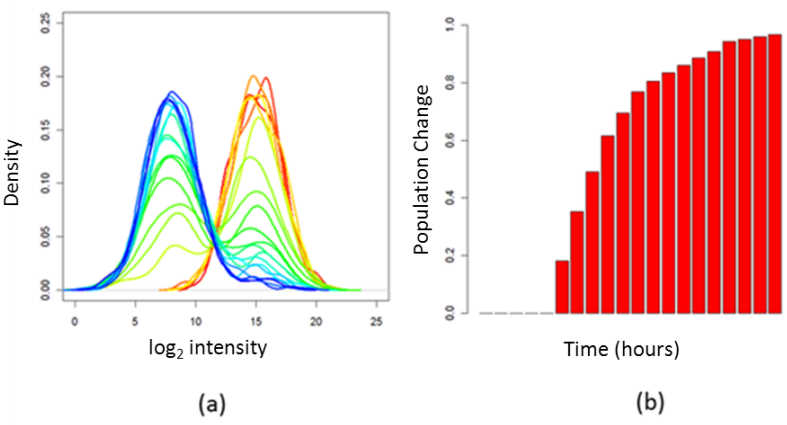


Fig. 44. Simulation results for the state transition model described in Fig. 43, with the parameters $N = 400, T = 20, \mu_1 = 15, \sigma_1 = 2, \mu_0 = 8, \sigma_0 = 2, t_0 = 5, c = 0.2$. (a) gene expression distributions at different times, color coded from red to blue; (b), corresponding population shifts. Note that the number of shifted cells is the highest right after the onset time, and decreases gradually with time.

to be transformed, and on average, $N \times c = 400 \times 0.2 = 80$ cells will switch to the low-expression state. In the next time point $t = 6$, only about $400 - 80 = 320$ are left and, as a result, at the same transition probability a smaller number of cells will be actually transformed to the low-expression state. As the pool of cells in a high-expression state goes down, the number of transformed cells will also decrease. Such behavior is not consistent with what we have observed in real experiments, where the number of transformed cells per unit time starts at a low number, goes up quickly, and finally returns to a low number (see Figs. 39(c) and 40 (b) for an example). Such mechanistic difference suggests that the proposed model in Fig. 44 is overly simplified and some of its assumptions may not be appropriate for real experiments.

3. Different Onset Times for Different Cells

As the discussion in the previous section suggests, the assumption about identical onset times for each cell in the cell line is too restrictive. Even in a homogeneous cell line, all cells will compete for resources, such as nutrients, oxygen, etc. Moreover, it is possible that different cells will stay at different stages of their cell cycle. Hence, the micro-environment will differ from cell to cell, resulting in different, readiness, or onset times. To account for such discrepancies, we consider a family of logistic growth curves that describe the onset times for individual cells. Consider the logistic function,

$$m(t) = \frac{K}{1 + e^{-r(t-t_1)}} \tag{4.3}$$

where t is time, $m = m(t)$ is the number of cells that have their cytoplasmic drug concentration above the threshold level (i.e., ready to be transformed cells), r and K are positive numbers, and t_1 is an arbitrary time. The logistic curve introduced in Eq. (4.3) is often used to model population dynamics in a resource limiting environment and has been applied in many fields including ecology, biology, etc. As we can see in Eq. (4.3), $\lim_{t\to\infty} m(t) = K$. Hence, K is also called the *carrying capacity*. t_1 is the inflection point of the logistic curve. It gives the time for $m(t)$ to reach the half height. The parameter r will affect the rate of increase.

There are several reasons to choose the logistic curve as a model for cell onset times. First, the carrying capacity K reflects the overall efficacy of the drug at a particular dosage. For example, for a population of 400 cells, a carrying capacity of 100 indicates that at maximum, a total of 100 cells will be eventually transformed. Second, the combined effect of r and t_1 will determine how soon the cells will be ready to be transformed, which is another important aspect of the drug effect. At higher dosage, we expect to the logistic curve to have a higher carrying capacity as

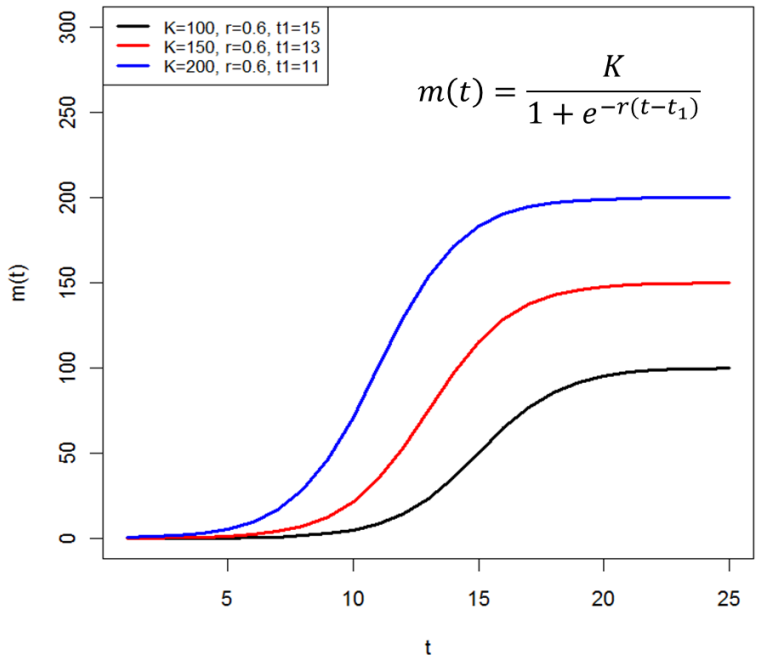


Fig. 45. Dosage dependent logistic curves to model cells' onset times. $m(t)$ is the number of cells that are ready to be transformed. Higher dosage should rise up earlier and have a higher carrying capacity K.

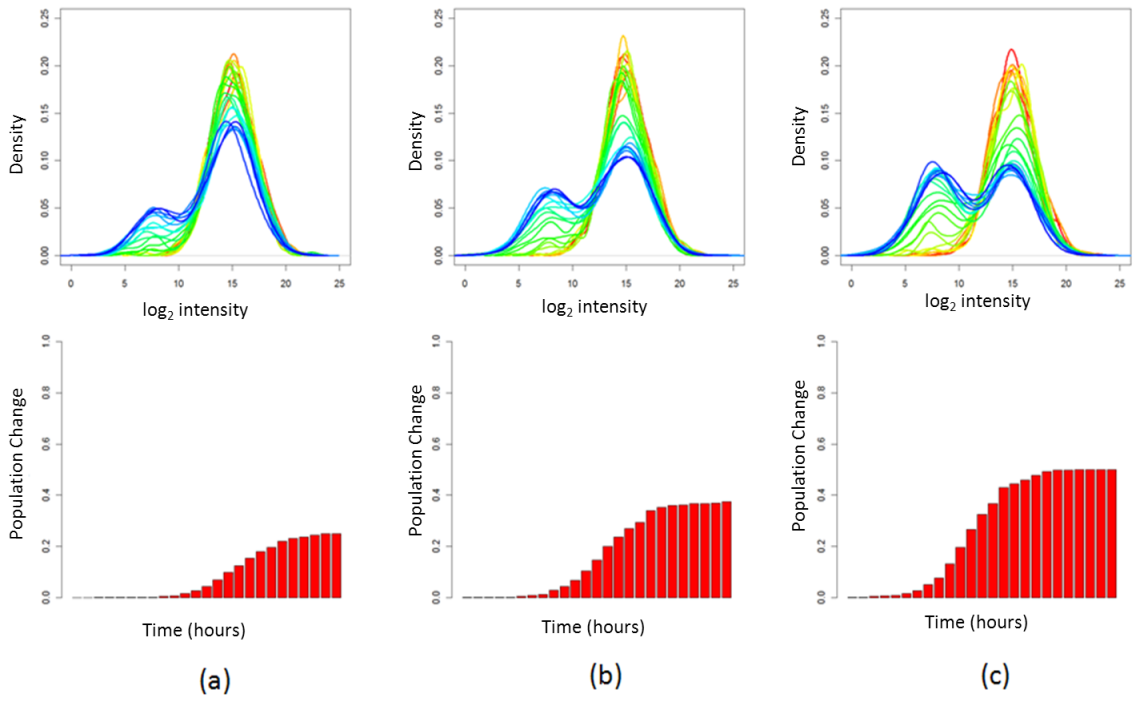


Fig. 46. Simulation results for cell population with different onset times, with $N = 400, T = 25, \mu_1 = 15, \sigma_1 = 2, \mu_0 = 8, \sigma_0 = 2, c = 0.5$. (a), (b), (c) corresponds to the logistic curves in Fig. 45 for the black, red and blue curves respectively.

well as an earlier rising period. Three hypothetical dosage dependent logistic curves are shown in Fig. 45, where the blue curve corresponds to the highest dosage level among all three.

Fig. 46 shows the simulation results for the three different cases in Fig. 46, with $N = 400, T = 25, \mu_1 = 15, \sigma_1 = 2, \mu_0 = 8, \sigma_0 = 2, c = 0.5$. The number of transformed cells per unit time starts low, goes higher with time, and eventually returns to 0. Such behavior is consistent with real experimental observations. Moreover, the final number of transformed cells is determined by the carrying capacity of the respective logistic curve.

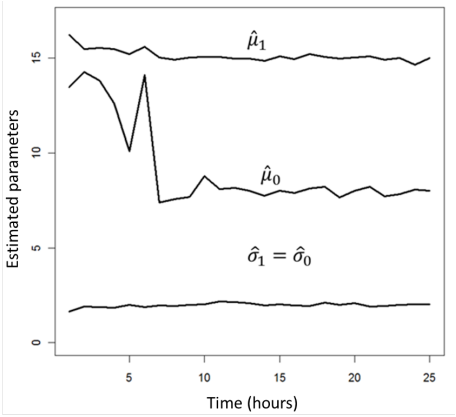


Fig. 47. Parameter estimation using EM algorithm for the expression distribution in Fig. 46(c). As we can see, except from the earlier time points, the estimated parameters agree very well with the true values $\mu_1 = 15, \sigma_1 = 2, \mu_0 = 8, \sigma_0 = 2$.

C. Model Parameter Estimation

1. Estimating Model Parameters: μ_1, σ_1, μ_0, σ_0

One can estimate the means and variances in the proposed model directly from the gene-expression distributions shown on Fig. 39(c). Given a mixture of Gaussian distributions, it is standard practice in statistics to use Expectation-Maximization (EM) algorithms to estimate the parameters of the individual components [75, 76]. Fig. 47 shows the EM algorithm estimation for the simulation results shown in Fig. 46(c), assuming equal variance of the individual components. Except for the earlier time points, the EM estimation is very close to the true value. We also tested the EM performance on real drug treatment experiment data (Fig. 48(a)). The results are shown on Fig. 48(b). Interestingly, the estimated mean and variance are quite flat, indicating that they might be time invariant.

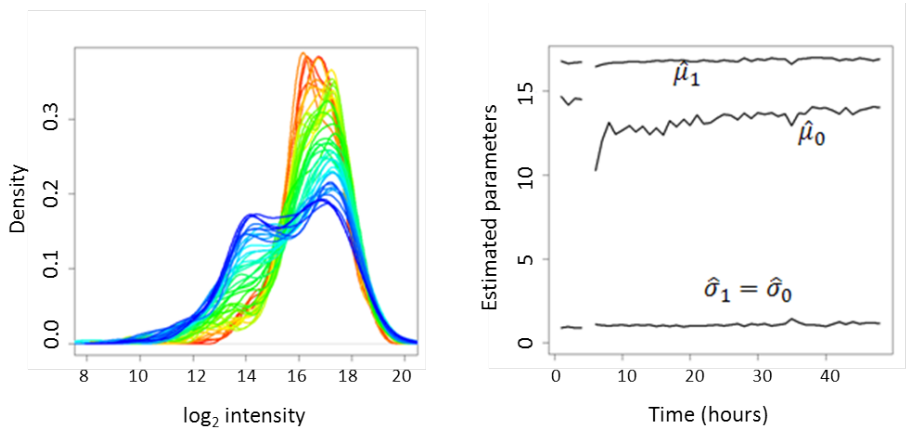


Fig. 48. Parameter estimation using EM algorithm for a real drug experiment (MKI67 responses to Lapatinib at dosage 2uM). The flatness of the estimated parameters indicates that they are time invariant, agreeing with our model assumptions.

2. Estimating Model Parameters: The Onset Time t_0

Estimation of the onset time is crucial to characterizing different dosage effects. As explained in Section B.3, the onset time for different cells is assumed to be governed by some logistic function. Hence, it suffices to estimate the three parameters in the logistic function.

For simplicity, let us assume that once a cell has reached its onset time t_0, it will immediately transform from the expression state 1 to the expression state 0. That is, we assume the transition probability $a_{10}^{i,1} = c$ in Fig. 43 (b) to be 1. In such a situation, the population shift dynamics is completely governed by the logistic curve, i.e., the number of actually transformed cells at time t is equal to the number, $m(t)$, of cells ready to be transformed. Hence, we want to find r and K and t_1 in Eq. (4.3) to minimize the square error [77]:

$$e = \sum_{t=1}^{T} (m(t) - n(t))^2 \tag{4.4}$$

where $m(t)$ is defined in Eq. (4.3) and $n(t)$ is the observed number of transformed cells at time t. To minimize the least square error in Eq. (4.4), we define $m(t) = Kh(t)$, where

$$h(t) = \frac{1}{1 + e^{(-r(t-t_1)}}$$

Hence, we can rewrite the error e to be

$$e = ||K\boldsymbol{H} - \boldsymbol{N}||^2 = K^2 < \boldsymbol{H}, \boldsymbol{H} > -2K < \boldsymbol{H}, \boldsymbol{N} > + < \boldsymbol{N}, \boldsymbol{N} > \tag{4.5}$$

where $\boldsymbol{H} = (h(t=1), h(t=2), \ldots, h(t=T))$, $\boldsymbol{N} = (n(t=1), n(t=2), \ldots, n(t=T))$, and $< \boldsymbol{X}, \boldsymbol{Y} >$ denotes the inner product. To minimize e, we set its partial derivatives with respective to K equal to 0, $\partial e / \partial K = 0$. Solving yields

$$K = \frac{< \boldsymbol{H}, \boldsymbol{N} >}{< \boldsymbol{H}, \boldsymbol{H} >} \tag{4.6}$$

Now, substitute this result into Eq. (4.5) to get

$$e =< \boldsymbol{N}, \boldsymbol{N} > - \frac{< \boldsymbol{H}, \boldsymbol{N} >^2}{< \boldsymbol{H}, \boldsymbol{H} >} \tag{4.7}$$

Eq. (4.7) contains just two parameters, r and t_1, the parameter K being eliminated. One can use gradient descent method to find r and t_1 to minimize the error. K can be computed from Eq. (4.6). Fig. 49 shows the results for fitting logistic curves on 4 different dose response population-shift data. The closeness of the fitted curve and the real data indicates that the proposed model for onset times provides a good approximation. Also notice that, at low dosage, doubling the concentration almost doubles K; however, at high dosage, doubling the concentration does not increase K significantly. This indicates that the drug has reached the saturating effect at around

16 uM.

D. Conclusion and Future Study

In this chapter, we propose a model to describe cell population gene-expression dynamics after drug treatment. Moreover, we show that under some simplified conditions, the model parameters can be inferred from data and the parameters bear useful biological implications, e.g. the carrying capacity K. Such model facilitates systematic understanding of drug response dynamics, and therefore is useful for drug development in general.

Along the same line of research, there are several interesting biological questions to be answered. First, what is the source of GFP intensity variation from cell to cell? As explained in the introduction section, the variation could be attributed to two reasons: either internally (transcription bursts) or externally (random GFP cassette insertion site). If it is due to the second reason, isolating a single cell from the cell population and repopulate a monoclonal cell line from the single cell should eliminate cell-to-cell GFP intensity variations, since the GFP insertion sites will be exactly the same for the derived monoclonal cell line. On the other hand, if the source of variation is due to transcription bursts, even the monoclonal cell line will still exhibit the same degree of variations. Second, what affects the carrying capacity or the final responded number of cells to a particular drug? As we have shown in the paper, the carrying capacity is dosage dependent, and there seems to be a dosage level above which the carrying capacity is saturated. Going a step deeper, why is this so? Is it because the cell line is heterogeneous and there is a subpopulation of cells that are immune to the drug? Or, is it because the drug has degraded after 48 hours and has lost its potency? If it is due to the second reason, we should get an increase of the

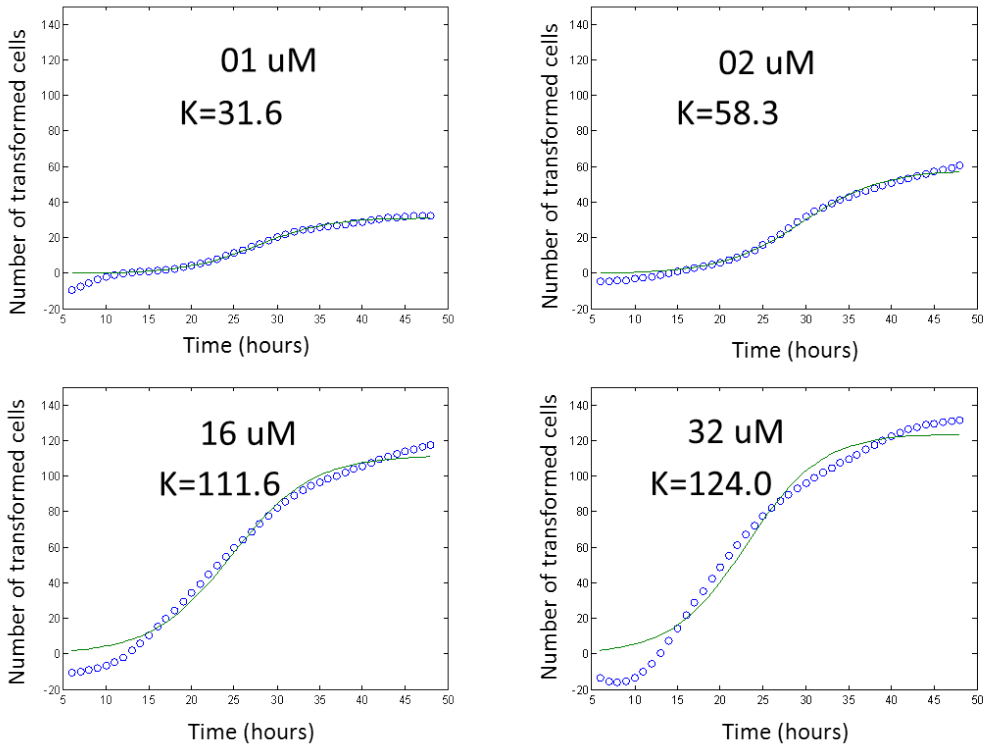


Fig. 49. Fitting logistic curves for population shift at different dosage levels. Circled line: observed number of transformed cell; solid line: fitted logistic curve. K is the estimated carrying capacity of the respective dosage. At low dosage, doubling the concentration almost doubles K, however, at high dosage, doubling the concentration does not increase K significantly. This indicates that the drug has reached the saturating effect at around 16 uM.

carrying capacity by replacing the cell media and refresh the drug after 48 hours. In the future study, we plan to carry out experiments to answer those questions.

CHAPTER V

CONCLUSION

In this dissertation, we have focused on applying engineering and statistical approaches to specific fields of cancer research with implications for designing drug intervention strategies. Here, we summarize the main contributions of this work.

- We have proposed a model based framework for the characterization and detection of master genes and canalizing genes. The framework can readily incorporate prior pathway knowledge, which is a unique feature that distinguishes us from the non-model based approaches proposed previously [10, 11]. We have also shown how the detectability of master genes or canalizing genes is affected by the various network structures and the associated parameters. The model based approach ultimately enables us to unify the two concepts: both master genes and canalizing genes tend to be the "hub" genes of the network; however, there are still subtle differences between the two: master genes measure the ability to control while canalizing genes measure the ability of taking over control.

- We have developed a time series alignment algorithm that can robustly and accurately identify mechanistic similarities in drug responses. The algorithm can facilitate large scale cancer drug MOA comparisons, which is a slow and labor consuming process in the existing approach [19].

- We have proposed a Markov model that describes the gene expression dynamics for a population of cells after drug treatment. Unlike the traditional method that treats gene expression as an averaged behavior of a large number of different cells, our approach models gene expression on the single cell level and therefore

can account for cell to cell variabilities. This finer level characterization has led to a deeper understanding of cell line based drug intervention strategies. Motivated by the proposed model, we have also proposed a real experiment that can uncover the source of GFP intensity variations from cell to cell.

Along the same line of research, there are several aspects that can be further developed. First, the tree structured Bayesian network employed in Chapter II can be extended to incorporate more complicated network structures. To do so, one needs to infer Bayesian networks that are consistent with a set of predefined biological constraints, i.e., the prior pathway knowledge. To our knowledge, such problem has not yet been investigated in the Bayesian network research community. Second, for the drug MOA study, our current approach focuses on the similarities on the individual GFP reporter level; however, it is desirable to obtain a higher level similarity representation by aggregating results from individual GFP reporters according to their functional groups. Moreover, it is interesting to see if combining two drugs will produce synergistic effect based on their individual MOAs, i.e., combinatorial drug therapies. More experiments are needed to answer questions like: Is there a subpopulation of cancer cells that are resistant to the applied drug? What are the growth rates for each subpopulation if the cell line is heterogeneous? What is the optimal dosage and frequency to apply a cancer drug? What are the critical missing pieces in the pathway wiring diagram that render the failure of the applied drug? We believe that all the aforementioned questions can be answered in a model based and hypotheses driven framework similar as the ones used throughout this dissertation. It is our hope that this work will generate enough interests and enthusiasms for others to tackle those problems.

REFERENCES

[1] S. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*. New York: Oxford University Press, 1993.

[2] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002.

[3] A. Datta, A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in Markovian genetic regulatory networks," *Machine Learning*, vol. 52, no. 1/2, pp. 169–191, 2003.

[4] R. Pal, A. Datta, and E. R. Dougherty, "Optimal infinite-horizon control for probabilistic Boolean networks," *IEEE Trans Signal Process*, vol. 54, no. 6, pp. 2375–2387, 2006.

[5] S. Marshall, L. Yu, Y. Xiao, and E. R. Dougherty, "Inference of a probabilistic Boolean network from a single observed temporal sequence," *EURASIP J Bioinform Syst Biol*, vol. 2007, pp. 32 454–32 454, 2007.

[6] I. Ivanov, R. Pal, and E. R. Dougherty, "Dynamics preserving size reduction mappings for probabilistic Boolean networks," *IEEE Trans Signal Process*, vol. 55, no. 5, pp. 2310–2322, 2007.

[7] N. Ghaffari, I. Ivanov, X. Qian, and E. R. Dougherty, "A cod-based reduction algorithm for designing stationary control policies on Boolean networks," *Bioinformatics*, vol. 26, no. 12, pp. 1556–1563, 2010.

[8] X. Qian, N. Ghaffari, I. Ivanov, and E. R. Dougherty, "State reduction for network intervention in probabilistic Boolean networks," *Bioinformatics*, vol. 26, no. 24, pp. 3098–3104, 2010.

[9] D. M. Chickering, D. Heckerman, C. Meek, and D. Madigan, "Learning Bayesian networks is NP-hard," Microsoft Research, Tech. Rep., 1994.

[10] E. R. Dougherty, M. Brun, J. M. Trent, and M. L. Bittner, "Conditioning-based modeling of contextual genomic regulation," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 6, no. 2, pp. 310–320, 2009.

[11] D. C. Martins, U. M. Braga-Neto, R. F. Hashimoto, M. L. Bittner, and E. R. Dougherty, "Intrinsically multivariate predictive genes," *IEEE J Sel Top Signal Process*, vol. 2, no. 3, pp. 424–439, 2008.

[12] M. S. Carro, W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch, H. Colman, A. Lasorella, K. Aldape, A. Califano, and A. Iavarone, "The transcriptional network for mesenchymal transformation of brain tumours," *Nature*, vol. 463, no. 7279, pp. 318–318, 2010.

[13] C. Lefebvre, P. Rajbhandari, M. J. Alvarez, P. Bandaru, W. K. Lim, M. Sato, K. Wang, P. Sumazin, M. Kustagi, B. C. Bisikirska, K. Basso, P. Beltrao, N. Krogan, J. Gautier, R. Dalla-Favera, and A. Califano, "A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers," *Mol Syst Biol*, vol. 6, pp. 377–377, 2010.

[14] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7 Suppl 1, p. S7, 2006.

[15] C. Sima and E. R. Dougherty, "What should be expected from feature selection in small-sample settings," *Bioinformatics*, vol. 22, no. 19, pp. 2430–2436, 2006.

[16] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?" *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.

[17] B. Hanczar, J. Hua, and E. R. Dougherty, "Decorrelation of the true and estimated classifier errors in high-dimensional settings," *EURASIP J. Bioinformatics Syst. Biol.*, vol. 2007, no. 2, pp. 2:1–2:12, 2007.

[18] C. Zhao, M. L. Bittner, R. S. Chapkin, and E. R. Dougherty, "Characterization of the effectiveness of reporting lists of small feature sets relative to the accuracy of the prior biological knowledge," *Cancer Inform*, vol. 9, pp. 49–60, 2010.

[19] J. Hua, C. Sima, M. Cypert, C. Gooden, S. Shack, L. Alla, E. Smith, J. M. Trent, E. R. Dougherty, and M. L. Bittner, "Tracking transcriptional activities with high-throughput epifluorescent imaging," *Submitted to the Journal of Biomedical Optics*, 2012.

[20] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Mol Biol Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[21] A. B. Khodursky, B. J. Peter, N. R. Cozzarelli, D. Botstein, P. O. Brown, and C. Yanofsky, "DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli," *Proc Natl Acad Sci*, vol. 97, no. 22, pp. 12 170–12 175, 2000.

[22] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[23] J. Aach and G. M. Church, "Aligning gene expression time series with time warping algorithms," *Bioinformatics*, vol. 17, no. 6, pp. 495–508, 2001.

[24] X. Liu and H. G. Mller, "Modes and clustering for time-warped gene expression profile data," *Bioinformatics*, vol. 19, no. 15, pp. 1937–1944, 2003.

[25] J. Criel and E. Tsiporkova, "Gene time echipression warper: a tool for alignment, template matching and visualization of gene expression time series," *Bioinformatics*, vol. 22, no. 2, pp. 251–252, 2006.

[26] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic gene expression in a single cell," *Science*, vol. 297, no. 5584, pp. 1183–1186, 2002.

[27] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, "Stochastic mrna synthesis in mammalian cells," *PLoS Biol*, vol. 4, no. 10, pp. e309–e309, 2006.

[28] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[29] S. R. Eddy, "Multiple alignment using hidden Markov models," *Proc Int Conf Intell Syst Mol Biol*, vol. 3, pp. 114–120, 1995.

[30] S. Eddy, "Profile hidden Markov models," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.

[31] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[32] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 379–385, 1992.

[33] M. Wang, X. Zhou, R. W. King, and S. T. Wong, "Context based mixture model for cell phase identification in automated fluorescence microscopy," *BMC Bioinformatics*, vol. 8, no. 1, pp. 32–32, 2007.

[34] J. Downward, "Targeting RAS signalling pathways in cancer therapy," *Nat. Rev. Cancer*, vol. 3, no. 1, pp. 11–22, 2003.

[35] R. A. Weinberg, *The Biology of Cancer*. New York: Garland Science, 2007.

[36] H. Davies, G. R. Bignell, C. Cox, P. Stephens, S. Edkins, S. Clegg, J. Teague, H. Woffendin, M. J. Garnett, W. Bottomley, N. Davis, E. Dicks, R. Ewing, Y. Floyd, K. Gray, S. Hall, R. Hawes, J. Hughes, V. Kosmidou, A. Menzies, C. Mould, A. Parker, C. Stevens, S. Watt, S. Hooper, R. Wilson, H. Jayatilake, B. A. Gusterson, C. Cooper, J. Shipley, D. Hargrave, K. Pritchard-Jones, N. Maitland, G. Chenevix-Trench, G. J. Riggins, D. D. Bigner, G. Palmieri, A. Cossu, A. Flanagan, A. Nicholson, J. W. Ho, S. Y. Leung, S. T. Yuen, B. L. Weber, H. F. Seigler, T. L. Darrow, H. Paterson, R. Marais, C. J. Marshall, R. Wooster, M. R. Stratton, and P. A. Futreal, "Mutations of the BRAF gene in human cancer," *Nature*, vol. 417, no. 6892, pp. 949–954, 2002.

[37] S. Imoto, Y. Tamada, C. J. Savoie, and S. Miyano, "Analysis of gene networks for

drug target discovery and validation," *Methods Mol. Biol.*, vol. 360, pp. 33–56, 2007.

[38] R. Layek, A. Datta, M. Bittner, and E. R. Dougherty, "Cancer therapy design based on pathway logic," *Bioinformatics*, vol. 27, no. 4, pp. 548–555, 2011.

[39] S. V. Sharma, D. A. Haber, and J. Settleman, "Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents," *Nat. Rev. Cancer*, vol. 10, pp. 241–253, 2010.

[40] C. H. Waddington, "Canalization of development and the inheritance of acquired characters," *Nature*, vol. 150, pp. 563–565, 1942.

[41] C. J. Tabin and R. A. Weinberg, "Analysis of viral and somatic activations of the cHa-ras gene," *J. Virol.*, vol. 53, pp. 260–265, 1985.

[42] M. Gomez-Lazaro, F. J. Fernandez-Gomez, and J. Jordan, "p53: twenty five years understanding the mechanism of genome protection," *J. Physiol. Biochem.*, vol. 60, no. 4, pp. 287–307, 2004.

[43] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann, 1988.

[44] F. V. Jensen, *Bayesian Networks and Decision Graphs*. Berlin, Germany: Springer, 2001.

[45] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: The MIT Press, 2009.

[46] R. E. Neapolitan, *Learning Bayesian Networks* . Upper Saddle River, NJ: Prentice Hall, 2003.

[47] J. Pearl, "Reverend Bayes on inference engines: A distributed hierarchical approach." *Proc. of the Second National Conference on Artificial Intelligence*, pp. 133–136, 1982.

[48] E. R. Dougherty, S. Kim, and Y. Chen, "Coeffcient of determination in nonlinear signal processing," *Signal Processing*, vol. 80, no. 10, pp. 2219–2235, 2000.

[49] S. Kim, E. R. Dougherty, Y. Chen, K. Sivakumar, P. Meltzer, J. M. Trent, and M. Bittner, "Multivariate measurement of gene expression relationships," *Genomics*, vol. 67, pp. 201–209, 2000.

[50] R. F. Hashimoto, S. Kim, I. Shmulevich, W. Zhang, M. L. Bittner, and E. R. Dougherty, "Growing genetic regulatory networks from seed genes," *Bioinformatics*, vol. 20, no. 8, pp. 1241–1247, 2004.

[51] R. S. Chapkin, C. Zhao, I. Ivanov, L. A. Davidson, J. S. Goldsby, J. R. Lupton, R. A. Mathai, M. H. Monaco, D. Rai, W. M. Russell, S. M. Donovan, and E. R. Dougherty, "Noninvasive stool-based detection of infant gastrointestinal development using gene expression profiles from exfoliated epithelial cells," *Am. J. Physiol. Gastrointest. Liver Physiol.*, vol. 298, no. 5, pp. G582–589, 2010.

[52] T. Chen and U. Braga-Neto, "Exact performance of CoD estimators in discrete prediction," *EURASIP Journal on Advances in Signal Processing*, no. 1, pp. 1–13, 2010.

[53] S. Kauffman, "Homeostasis and differentiation in random genetic control networks," *Nature*, vol. 224, pp. 177–178, 1969.

[54] I. Shmulevich, H. Lahdesmaki, E. R. Dougherty, J. Astola, and W. Zhang, "The role of certain post classes in Boolean network models of genetic networks," *Proc.*

Natl. Acad. Sci., vol. 100, no. 19, pp. 10 734–10 739, 2003.

[55] I. Shmulevich and S. A. Kauffman, "Activities and sensitivities in Boolean network models," *Phys. Rev. Lett.*, vol. 93, no. 4, p. 048701, 2004.

[56] S. Harris, B. Sawhill, A. Wuensche, and S. Kauffman, "A model of transcriptional regulatory networks based on biases in the observed regulation rules," *COMPLEXITY*, vol. 7, pp. 23–40, 2002.

[57] E. R. Dougherty, *Probability and Statistics for the Engineering, Computing and Physical Sciences*. Upper Saddle River, NJ: Prentice Hall, 1990.

[58] J. Pearl, "Fusion, propagation, and structuring in belief networks," *Artificial Intelligence (Elsevier)*, vol. 29, no. 3, pp. 241–288, 1986.

[59] G. Rebane and J. Pearl, "The recovery of causal poly-trees from statistical data," *Proceedings of UAI*, pp. 222–228, 1987.

[60] U. Scherf, D. T. Ross, M. Waltham, L. H. Smith, J. K. Lee, L. Tanabe, K. W. Kohn, W. C. Reinhold, T. G. Myers, D. T. Andrews, D. A. Scudiero, M. B. Eisen, E. A. Sausville, Y. Pommier, D. Botstein, P. O. Brown, and J. N. Weinstein, "A gene expression database for the molecular pharmacology of cancer," *Nat Genet*, vol. 24, no. 3, pp. 236–44, 2000.

[61] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, and A. J. Butte, "Discovery and preclinical validation of drug indications using compendia of public gene expression data," *Sci Transl Med*, vol. 3, no. 96, p. 96ra77, 2011.

[62] S. J. Chen, Y. J. Zhu, J. H. Tong, S. Dong, W. Huang, Y. Chen, W. M. Xiang, L. Zhang, X. S. Li, and G. Q. Qian, "Rearrangements in the second intron of the

RARA gene are present in a large majority of patients with acute promyelocytic leukemia and are used as molecular marker for retinoic acid-induced leukemic cell differentiation," *Blood*, vol. 78, no. 10, pp. 2696–2701, 1991.

[63] N. V. Sergina, M. Rausch, D. Wang, J. Blair, B. Hann, K. M. Shokat, and M. M. Moasser, "Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3," *Nature*, vol. 445, no. 7126, pp. 437–41, 2007.

[64] R. Z. Yusuf, Z. Duan, D. E. Lamendola, R. T. Penson, and M. V. Seiden, "Paclitaxel resistance: molecular mechanisms and pharmacologic manipulation," *Curr Cancer Drug Targets*, vol. 3, no. 1, pp. 1–19, 2003.

[65] M. Chalfie, Y. Tu, G. Euskirchen, W. W. Ward, and D. C. Prasher, "Green fluorescent protein as a marker for gene expression," *Science*, vol. 263, no. 5148, pp. 802–805, 1994.

[66] R. Hunt-Newbury, R. Viveiros, R. Johnsen, A. Mah, D. Anastas, L. Fang, E. Halfnight, D. Lee, J. Lin, A. Lorch, S. McKay, H. M. Okada, J. Pan, A. K. Schulz, D. Tu, K. Wong, Z. Zhao, A. Alexeyenko, T. Burglin, E. Sonnhammer, R. Schnabel, S. J. Jones, M. A. Marra, D. L. Baillie, and D. G. Moerman, "High-throughput in vivo analysis of gene expression in Caenorhabditis elegans," *PLoS Biol*, vol. 5, no. 9, p. e237, 2007.

[67] E. R. Dougherty and R. A. Lotufo, *Hands-on Morphological Image Processing*. Bellingham, WA: SPIE Optical Engineering Press, 2003.

[68] P. Tormene, T. Giorgino, S. Quaglini, and M. Stefanelli, "Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation," *Artif Intell Med*, vol. 45, no. 1, pp. 11–34, 2009.

[69] D. S. Hirschberg, "Algorithms for the longest common subsequence problem," *J. ACM*, vol. 24, no. 4, pp. 664–675, 1977.

[70] L. Bergroth, H. Hakonen, and T. Raita, "A survey of longest common subsequence algorithms," in *Seventh International Symposium on String Processing and Information Retrieval*, 2000, pp. 39–48.

[71] J. Gerdes, "Ki-67 and other proliferation markers useful for immunohistological diagnostic and prognostic evaluations in human malignancies," *Semin Cancer Biol*, vol. 1, no. 3, pp. 199–206, 1990.

[72] T. Scholzen and J. Gerdes, "The Ki-67 protein: from the known and the unknown," *J Cell Physiol*, vol. 182, no. 3, pp. 311–22, 2000.

[73] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, "Real-time kinetics of gene activity in individual bacteria," *Cell*, vol. 123, no. 6, pp. 1025–36, 2005.

[74] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, "Stochastic mrna synthesis in mammalian cells," *PLoS Biol*, vol. 4, no. 10, p. e309, 2006.

[75] T. K. Moon, "The expectation-maximization algorithm," *Signal Processing Magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.

[76] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden Markov models," University of California Berkeley, Tech. Rep., 1998.

[77] F. Cavallini, "Fitting a logistic curve to data," *College Mathematics Journal*, vol. 24, no. 3, pp. 247–253, 1993.

VITA

Chen Zhao received his B.S. degree in telecommunication engineering from Beijing University of Posts and Telecommunications, Beijing, China, in July 2006, and his M.Eng. degree in the Department of Electrical and Computer Engineering from Texas A&M University, College Station, TX, USA, in December 2007, focusing on magnetic resonance imaging. He has been pursuing his Ph.D. degree in electrical and computer engineering at Texas A&M University since June 2008 and defended his Ph.D. in December 2011. During his Ph.D., he worked as a research assistant in the Genomic Signal Process (GSP) lab, focusing on statistical pattern recognition applications to biomarker discovery as well as pathway based drug response data analysis and modeling. During January–December 2011, he was a research intern at the Center for Proteomics at the Translational Genomics Research Institute, Phoenix, AZ, where he developed proteomics skills. He can be reached at tamuzc0611@gmail.com or department of electrical and computer engineering, c/o Dr. Edward R. Dougherty, Texas A&M University, College Station, TX, 77843-3128.

The typist for this dissertation was Chen Zhao.